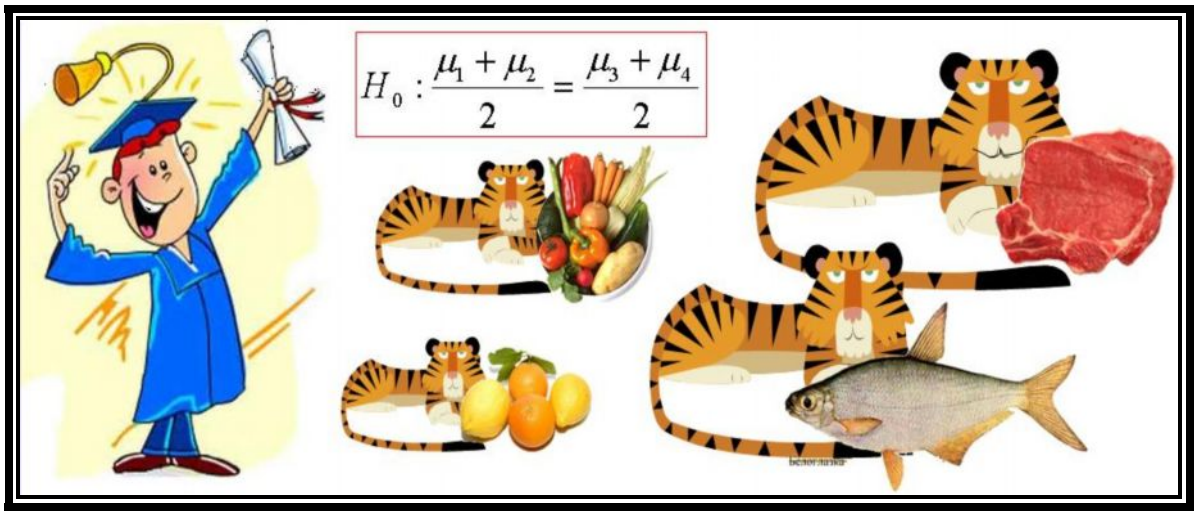
Российская академия наук

Институт экологии Волжского бассейна

В.К. Шитиков, Г.С. Розенберг

Рандомизация и бутстреп:

статистический анализ в биологии и экологии с использованием R



Исправленная и дополненная интернет-версия от 15.11.2013

Тольятти 2013

Шитиков В.К., Розенберг Г.С. Рандомизация и бутстреп: статистический анализ в биологии и экологии с использованием R. – Тольятти: Кассандра, 2013. – 314 с.

ISBN

В книге представлено описание широкой панорамы статистических методов, как повсеместно используемых, так и не нашедших пока должного применения в обработке данных экологического мониторинга. Сюда вошли элементарная статистика, проверка гипотез, различные подходы к оценке биоразнообразия, дисперсионный анализ, специальные формы регрессии и оценки информативного набора предикторов моделей, многомерные методы классификации, редукции и распознавания образов, процедуры, использующие байесовский подход, анализ временной или пространственной динамики и т.д. Мы не ставили целью подробно описать теоретические аспекты всех этих методов, но широко иллюстрировали методику их применения на примерах биологического характера.

Совокупность представленных методов связывается двумя основополагающими идеями. Во-первых, в каждом примере мы пытались найти "изюминку" в виде использования нового класса компьютерно-интенсивных (computer-intensive) методов, в широком смысле относящихся к семейству различных процедур Монте-Карло. Наиболее детально представлен численный ресамплинг, который заключается в различных технологиях генерации повторных выборок. Описаны алгоритмы, включающие рандомизацию, перестановочный тест (permutation), бутстреп (bootstrap), метод "складного ножа" (jackknife) и кросс-проверку (cross-validation). Мы показываем, как с их помощью можно корректно проверить статистическую гипотезу или получить несмещенные характеристики искомого параметра: оценки математического ожидания, дисперсии, доверительного интервала, коэффициентов модели. Где это возможно, мы сравниваем полученные результаты с классическими асимптотическими методами, использующими то или иное стандартное предельное распределение.

Вторая "красная нить" - возможность для читателей легко воспроизвести самим технику выполнения расчетов. Мы ориентировались на статистическую среду R, которая постепенно становится общепризнанным мировым стандартом при проведении научно-технических расчетов. В конце каждого раздела нами представлены тексты несложных скриптов в кодах R, позволяющих выполнить самостоятельно статистический анализ рассматриваемых примеров. В этой связи, представляемая монография может рассматриваться также как справочник по реализации различных алгоритмов обработки данных для исследователей, которых привлекла эта инструментальная среда.

Книга может быть использована в качестве учебного пособия по статистическим методам для студентов и аспирантов высших учебных заведений биологического профиля.

Табл. 40, ил. 131. Библиогр. 232 назв.

Рецензент: д.б.н., профессор А.А. Савельев (г. Казань)
Рекомендовано к печати Ученым советом Института экологии Волжского бассейна РАН (протокол № 11 от 22 октября 2013 г.).

445003, Россия, Самарская обл., г. Тольятти, ул. Комзина, 10
Институт экологии Волжского бассейна РАН
Тел., факс: (8482) 489-504; E-mail: ievbras2005@mail.ru
Сайт авторов: http://www.ievbras.ru/ecostat/Kiril

СОДЕРЖАНИЕ

Четырем читателям – вы знаете, кто вы…
Дж. Аберкромби

ПРЕДИСЛОВИЕ

Стремительное изменение современного мира, связанное с революционными достижениями вычислительной техники, информационных технологий и связи, обеспечило возможность быстрого, комплексного и точного анализа больших массивов данных. Высокопроизводительные компьютеры и общедоступное программное обеспечение позволили представлять результаты графически и в понятной информативной форме способами, ранее недоступными с помощью ручки и бумаги.

Менее очевидный процесс связан с коренным пересмотром основных концепций прикладной статистики. В докомпьютерный период, когда обработка данных требовала много времени и усилий, делался акцент на методы, которые позволили бы получить максимум информации при небольшом объеме вычислений. Общий подход был весьма прост: делалось предположение, что структура полученных данных “похожа” на некоторую распространенную статистическую модель (например, подчиняется нормальному распределению), после чего выборочные оценки параметров рассчитывались по относительно простым теоретическим формулам.

Однако для сложных систем (прежде всего, экономических и экологических), которые рассматриваются как статистические ансамбли, состоящие из большого количества неоднородных компонент, в структуре данных наблюдается существенное отличие от обычных *гауссовых* распределений. В частности, феномен негауссовости заключается в том, что в результате увеличения объема выборки некоторые оцениваемые параметры генеральной совокупности (в первую очередь, дисперсия) начинают монотонно возрастать, т.е. данные перестают подчиняться центральной предельной теореме теории вероятностей (Хайтун, 1983). В этих случаях выводы, основанные на предположениях о нормальности, часто не являются корректными и поэтому практически оказываются не всегда полезными.

Появление компьютеров в корне изменило концепцию обработки данных, так как вычисления стали быстры и необременительны, а во краю угла встало требование корректности формируемых выводов. Известный американский статистик, профессор Станфордского университета Б. Эфрон написал статью под названием «Компьютеры и статистика: подумаем о невероятном» (Efron, 1979a), в которой обосновал развитие нового класса альтернативных компьютерно-интенсивных (computer-intensive) технологий, включающих рандомизацию, бутстреп и методы Монте-Карло. Эти технологии, объединенные общим термином "*численный ресамплинг*", не требуют никакой априорной информации о законе распределения изучаемой случайной величины. Вместо этого они выполняют многократную обработку различных фрагментов исходного массива эмпирических данных, как бы рассматривая их под различными углами зрения и сопоставляя полученные таким образом результаты.

С учетом этого можно предположить, что развитие прикладной статистики пойдет по двум различным путям. Первый заключается в развитии традиционного “асимптотического” направления и в его рамках расширяется арсенал методик и новых критериев, которые могут оказаться более предпочтительными в тех или иных условиях обработки данных. Но, например, в ходе дисперсионного анализа при различных его модификациях рекомендовано к использованию около трех десятков “именных” критериев (Дана, Коновера, Джонкхиера-Терпстра, Бартлетта, Кокрена, Шеффе, Дункана, Тьюки, Левене, Брауна-Форсайта, Бхапкара, Дешпанде, Краскела–Уоллиса, Фридмана, Квейда, Пэйджа, Хотеллинга, Джеймса-Сю, Пури-Сена-Тамура, Шейрера-Рэя-Хэйра, Уилкса, Кульбака и др.), для проверки нормальности распределения – более двух десятков критериев согласия, а в непараметрической статистике число методик сравнения выборок,

представленных в справочниках (Гайдышев, 2001; Кобзарь, 2006), приближается к сорока. Области использования каждого из этих вариантов выглядят размытыми, а отмечаемые достоинства и недостатки субъективны и противоречивы, что часто приводит в растерянность конечных пользователей. Альтернативный путь сводится к разработке методически единых универсальных алгоритмов поиска решения (например, формирования частотного распределения анализируемого показателя в результате многократных итераций). Это позволяет только за счет интенсивной работы компьютера провести надежное тестирование данных без строгой привязки к формуле применяемого критерия. Так как статистика неизбежно основана на вычислениях, эффективность и результативность их реализации должна быть наиболее важным и объективным аргументом в решении, какой из этих двух путей обработки данных лучше подходит для широкого круга прикладных задач.

Ресамплинг основывается на традиционных общих идеях статистического анализа. Фундаментальным остается рассуждение о соотношении между случайными повторностями эмпирических данных и генеральной совокупностью, причем никакие сверхинтенсивные методы не являются панацеей от влияния неучтенных факторов или систематических погрешностей при плохо поставленном эксперименте. Статистические выводы также базируются на классических доверительных интервалах и *p*-значениях, основанных на выборочных распределениях используемых критериев. Ключевое отличие лишь в том, что повторности классической выборки извлекаются из генеральной совокупности, а псевдоповторности ресамплинга – из самой эмпирической выборки («The population is to the sample as the sample is to the bootstrap samples» – Fox, 2002).

При этом, во-первых, новые методы ресамплинга освобождают нас от необходимости делать не всегда обоснованные предположения типа «Пусть моделируемая случайная величина распределена по нормальному закону $N\left(m,\sigma^2\right)$». Во-вторых, они непосредственно обращаются к самой сути статистического анализа и показывают, как изменится распределение выборочных характеристик, если будет использовано практически неограниченное количество повторностей данных, полученных в тех же самых условиях. В-третьих, при достаточном количестве проведенных итераций методы ресамплинга дают более точные результаты, чем традиционные методы. Наконец, они концептуально более просты и освобождают нас от необходимости искать в справочниках различные математические формулы критериев, наиболее подходящих в конкретных условиях, и способов их аппроксимации.

В долгосрочной перспективе перечисленные преимущества ресамплинга могут оказать огромное влияние на стиль, с которым предмет статистики преподается и осуществляется на практике. Этому способствует издание и переиздание за рубежом фундаментальных работ и практических руководств в этом направлении, подготовленных Б.Эфроном, Р.Тибширани, Э.Эджингтоном, Дж. Саймоном, М.Черником, К.Луннеборгом, Р.Дэвисоном, Д.Хинкли и другими (Эфрон, 1988; Efron B., Tibshirani, 1993; Edgington, 1995; Simon, 1997; Chernick, 1999; Lunneborg, 2000; Davison, Hinkley, 2006).

К сожалению, русскоязычному читателю трудно встретить публикации, посвященные этой динамично развивающейся идеологии, поэтому в нашей монографии мы приводим краткое изложение сути двух основных процедур ресамплинга – рандомизации и бутстрепа – и иллюстрируем возможности применения компьютерно-ориентированных методов на широком круге конкретных примеров. При компоновке глав книги мы во многом ориентировались на рубрикатор известного пособия Б. Манли «Рандомизация, бутстреп и методы Монте-Карло в биологии» (Manly, 2007), однако часто наши корабли следовали разным курсом, а подготовку всех примеров, расчеты и их интерпретацию мы выполнили самостоятельно. Многие идеи, показавшиеся нам ценными, мы почерпнули из учебных материалов, представленных на сайте Университета в Вермонте проф. Д. Ховелом (Howell, 2009, 2010). Важным материалом для нас явились также серия практических руководств Ф. Гуда (Good, 2005a, 2005б, 2006), а также

классическая монография П. и Л. Лежандров «Количественная экология» (Legendre, Legendre, 1998), содержащая прекрасное описание многомерных методов.

Хотя сами процедуры рандомизации и бутстрепа концептуально очень просты, главная причина недостаточного практического использования этих методик расчета обычно объясняется отсутствием под руками необходимого программного обеспечения. В общем случае для реализации такой возможности имеются два подхода:

- использование программ, управляемых с помощью меню, таких как SPSS или чрезвычайно, хотя и незаслуженно популярный пакет Statistica (в котором нам так и не удалось найти внятных возможностей использовать бутстреп-процедур во многих важных пунктах анализа данных);
- запись макроопределений на высокоуровневых интерпретируемых языках в вычислительных интерактивных средах, таких как R, MatLab, SAS, Stats, или самостоятельная разработка программ с использованием Visual Basic, C++, Delphi и др.

Принимая во внимание, что большинство пользователей для реализации расчетов оказывают предпочтение программам, управляемым с помощью меню, мы для нескольких примеров в начальных главах постарались использовать компактные версии удобных, бесплатных и "биологически ориентированных" компьютерных программ, созданных усилиями следующих авторов:

- Д. Ховела (D. Howell), программа «Resampling», представлена на сайте http://www.uvm.edu/~dhowell/StatPages/Resampling/Resampling.html ;
- П. Ядвижчака (P.Jadwiszczack), RundomPro 3.14, http://pjadw.tripod.com ;
- Б. Манли (B. Manly), пакет RT (randomization testing) из 11 программ, http://www.west-inc.com/computerprograms.html ;
- Ф. Хаммера и др. (Ø. Hammer, D. Harper, P. Ryan), разработавших набор разнообразных статистических функций для анализа палеонтологических данных **PAST**, где активно представлено уточнение выборочных параметров бутстреп-методом, http://folk.uio.no/ohammer/past и некоторых других.

Однако очень скоро обнаружилось, что возможностей этих программ будет читателю явно недостаточно, что дало нам повод обратиться к использованию статистической среды R (www.r-project.org), которая постепенно становится общепризнанным "стандартом" при проведении научно-технических расчетов в большинстве западных университетских центров и многих ведущих фирмах. Язык вычислений R, хотя и требует определенных усилий для своего освоения, зато позволяет оперативно выполнить расчеты, по своему разнообразию практически «столь же неисчерпаемые, как атом». Эта безграничность опирается на мощный "коллективный разум" тысяч бескорыстных разработчиков-интеллектуалов, среди которых нам хотелось бы отметить Б. Рипли (Ripley), К. Халворсена (Halvorsen), А. Канти (Canty) и других создателей пакетов для реализации ресамплинга, а также Я.Оксанена (Oksanen), Р. Киндта (Kindt), Х. Стивенса (Stevens), С. Госли (Goslee), Д. Урбана (Urban), С. Дрея (Dray) и А. Дюфора (Dufour), активно внедривших этот подход в экологические приложения.

Вопреки устоявшейся традиции зарубежных изданий подобного типа, мы не стали приводить развернутого описания языка R и его стандартных функций, поскольку оно уже достаточно полно представлено в документации и многочисленных учебных пособиях (Зарядов, 2010а, 2010б; Шипунов и др., 2012; Venables, Ripley, 2003; Crawley, 2007; Logan, 2010; Borcard et al., 2011). Отметим, что особым источником "творческого вдохновения" послужили для нас также обширные материалы по использованию среды R для обучения студентов экологических специальностей университетов в Монреале (проф. П. Лежандр – Legendre) и Монтане (проф. Д. Робертс – Roberts). Ссылки на наиболее интересные сайты для самообучения программированию в R приведены в приложении 2.

Специалисты насчитывают к настоящему времени сотни тысяч изданных книг по математической и прикладной статистике (обычный читатель вряд ли в состоянии одолеть хотя бы 2% из них). Было бы бессмысленным писать очередное руководство, поэтому

первоначально мы не планировали останавливаться на теоретических аспектах используемых статистических методов, полагая, что читатель либо знаком с их основами, либо может легко получить необходимые знания из доступных источников. Однако, пойдя навстречу искренним пожеланиям наших рецензентов сделать книгу полнее и строже, мы включили в ее текст несколько отвлеченных рассуждений о природе анализируемых случайных величин, методах оценивания параметров, концепциях проверки статистических гипотез и др.

И все же основной акцент сделан на разборе нескольких десятков примеров в следующей последовательности: краткая содержательная постановка задачи, смысл алгоритма обработки данных с некоторыми расчетными формулами, основные полученные результаты и их возможная интерпретация. Не всегда методы ресамплинга и Монте-Карло составляли сущность обработки данных, но, где это было возможно, мы старались найти “рациональное зерно” их использования.

Для статистического анализа большинства рассматриваемых примеров нами были составлены тексты несложных скриптов в кодах R, которые представлены в конце каждого раздела. В этой связи, представляемая монография может позиционироваться также как справочник по реализации различных статистических методов в среде R. Подробно комментируемые скрипты, а также файлы с исходными данными, позволяющие выполнить расчеты по примерам, представленным в основных разделах, могут быть загружены из ресурса Интернет: http://www.ievbras.ru/ecostat/Kiril/Article/A32/Data.zip

Мы адресуем нашу книжку массовому слою студентов, аспирантов и молодых ученых – биологам, экологам, медикам. При этом ставим вполне конкретные задачи:

- разъясняем смысл и механизмы реализации методов ресамплинга для тех, что не имел них четкого представления и делаем это на уровне рекламных проспектов и простейших “картинок” Д. Ховела;
- одновременно, для тех, кто этим не удовлетворился, существенно расширяем спектр рассуждений о применимости ресамплинга на более сложные случаи: многофакторный ANOVA, регрессию, классификацию, ординацию (в меру наших скромных усилий);
- привлекаем внимание к использованию статистической среды R, если готовых инструментальных средств для проведения расчетов не хватает (наши примеры включают как простой запуск функций в одну строку для начинающих, так и более сложные языковые конструкции для "продвинутых");
- напоминаем о существовании таких прекрасных, но недостаточно обсуждаемых способов обработки данных, как кросс-проверка, генетический алгоритм, метод опорных векторов, тест Мантеля, случайный зонд Пиелу, дисперсионный анализ матриц дистанций М.Андерсона, бутстрепирование деревьев классификации и др.

В книге для иллюстрации методов мы отказались от использования имитируемых выборок или специально подобранных файлов типа “ирис Фишера”, а привели только “живые” примеры, максимально приближенные к “боевым” условиям. Не всегда при расчетах были получены результаты, интересные с экологической точки зрения, но наша цель – описать методику, а неудовлетворенные оппоненты могут поискать сами более качественные примеры для ее реализации.

Исходные данные для примеров составили в основном экспедиционные материалы Института экологии Волжского бассейна РАН, которые были любезно предоставлены сотрудниками лабораторий и их руководителями Т.Д. Зинченко, В.Б. Голубом, О.А. Розенцвет, А.Л. Маленевым, Г.В. Еплановой и другими, кому мы приносим свою глубокую благодарность. Полный указатель использованных примеров с кратким описанием и ссылками на некоторые публикации представлен в приложении 1.

Мы также выражаем искреннюю признательность проф. А.В. Коросову (Петрозаводский государственный университет), к.б.н. С.С. Крамаренко (Николаевский государственный аграрный университет [Украина]), к.б.н. В.Н. Якимову (Нижегородский университет им. Н.И. Лобачевского), к.б.н. Н.А. Чижиковой (Казанский государственный

университет), к.б.н. Д.Ю. Нохрину (Уральский филиал ВНИИВСГЭ РАСХН), к.м.н. С.В. Петрову (Гродненский государственный университет [Беларусь]), к.б.н. Рапопорт И.Б. (Институт экологии горных территорий КБНЦ РАН), Д. Зелены (D. Zelený, Университет Масарика, Брно [Чехия]), Л. Чавесу (L. Chaves, университет Хоккайдо [Япония]) и многим другим коллегам, которые высказали ценные замечания при обсуждении отдельных методов и разделов этой книги.

Особую благодарность мы высказываем рецензенту проф. А.А. Савельеву (Казанский государственный университет) и Н.А. Цейтлину (Natan Zeitlin, Геттинген [Германия]), которые терпеливо направляли нас на путь тщательно выверенного освещения статистического материала и использования устоявшихся “политкорректных” формулировок. Не всегда их усилия оказались успешными, поэтому все “огрехи”, которые, несомненно, найдет придирчивый читатель, следует отнести только в наш адрес.

Нельзя не сказать также слова благодарности некоторым фондам и программам, которые в той или иной степени способствовали появлению этой работы: мы благодарны Российскому гуманитарному научному фонду «Волжские земли в истории и культуре России» (грант 12-12-63005), Программе грантов Президента Российской Федерации для государственной поддержки ведущих научных школ Российской Федерации (грант НШ-3018.2012.4), Российскому фонду фундаментальных исследований РФФИ-Поволжье (грант 13-04-97004), программе фундаментальных исследований Президиума РАН «Живая природа: современное состояние и проблемы развития» и программе Отделения биологических наук РАН «Биологические ресурсы России: динамика в условиях глобальных климатических и антропогенных воздействий».

1. БУТСТРЕП И СТАТИСТИЧЕСКОЕ ОЦЕНИВАНИЕ ВЫБОРОЧНЫХ ХАРАКТЕРИСТИК

1.1. Оценка параметров: точечные и интервальные характеристики

Традиционно выделяют две основных задачи прикладной статистики: оценку параметров и проверку гипотез. Если проверка статистических гипотез является важнейшим и практически отработанным инструментом научного познания, то процедура оценки неизвестных параметров распределения изучаемой генеральной совокупности носит в некотором смысле предположительный и казуальный характер, не вполне поддающийся обобщению. В итоге классическая схема оценивания параметров в версии, изложенной в авторитетных учебниках по статистике, выглядит следующим образом:

1. Пусть результаты наблюдения являются измерениями некоторых признаков изучаемого объекта в форме чисел или иных свойств нечисловой природы. *Эмпирическая совокупность* значений изучаемой случайной величины[1] представлена при этом простой выборкой X_n объемом n или несколькими ее независимыми повторностями.

2. Случайность и независимость выборок определяются той вероятностной моделью, с помощью которой был задан способ или механизм порождения данных. Например, при биологическом мониторинге часто реализуется простой случайный выборочный процесс, когда из воображаемой *генеральной совокупности* (популяции большого объема) случайно извлекается качественно однородный набор изучаемых эмпирических объектов. При этом каждая из возможных комбинаций объектов имеет равную вероятность быть отобранной.

3. Генеральная совокупность – это в некотором смысле бесконечное множество значений признака X_∞ (т.е. при $n \to \infty$), имеющее в условиях эксперимента некоторое *распределение*, под которым понимается функция $\boldsymbol{F}$, связывающая значения случайной величины с вероятностью их появления в совокупности. Выборочный ряд наблюдений X_n является частной случайной реализацией вероятностного процесса (т.е. произвольным подвектором из X_∞) и имеет некоторое эмпирическое распределение $\boldsymbol{F}_n^*$. Поскольку точный вид закона распределения популяции в общем случае неизвестен, основные статистические выводы делаются в предположении, что при увеличении повторностей наблюдений их совокупное слаженное распределение асимптотически приближается к $\boldsymbol{F}$.

4. Наибольшая информация о генеральной совокупности концентрируется в наборе основных *выборочных характеристик*, т.е. показателей, описывающих положение, рассеяние данных или форму кривой плотности распределения эмпирических частот. В роли таких характеристик могут выступать любые математически измеримые функционалы от результатов наблюдений, используемые для получения статистических выводов (они также часто называются *статистиками*). Статистика, как функция случайной выборки, также является случайной величиной. Распределение выборочной совокупности обычно предполагается унимодальным и тогда примерами основных статистик являются:

- показатели *положения* или центральной тенденции: различные варианты оценки среднего значения (в том числе, среднее по Колмогорову, являющееся их обобщением), стандартная ошибка средних, медиана, мода;
- показатели *разброса* (рассеяния, масштаба): дисперсия, стандартное отклонение, среднее отклонение, размах, коэффициент вариации, средняя разность Джини,

---

[1] Случайная величина – это не последовательность чисел, выбранных наугад, например, с помощью псевдогенератора случайных чисел. Значения случайной переменной отбираются на основе изучаемого вероятностного процесса; просто ее наблюдаемые реализации не могут быть известны прежде, чем сделан эксперимент.

Здесь и далее *эксперимент* понимается в широком смысле как «исследование любых пространственно-временных процессов в мире» (Kempthome (1979, p. 124), т.е. технических, социальных, экономических, экологических и др.

квартили, межквартильный размах;

- показатели *формы* и сдвига распределения: коэффициент асимметрии, эксцесс и др.

Для выборок случайных величин нечисловой природы могут быть рассчитаны доля, ошибка доли и дисперсия доли.

5. Хотя результаты отдельных наблюдений могут иметь существенную изменчивость, выборочные характеристики, основанные на начальных и центральных моментах (среднее, дисперсия, асимметрия) или некоторых функций от них, часто обнаруживают замечательную устойчивость. Общий характер случайного варьирования этих статистик приобретает асимптотически-нормальный характер, а в пределе при $n \to \infty$ их случайная величина вырождается в неслучайную независимо от специфики генеральной совокупности. Таким образом, основные выборочные характеристики с ростом объема выборки стремятся к своим теоретическим аналогам и ведут себя как нормально распределенные случайные величины (Айвазян, Мхитарян, 1998).

Одна из главных целей статистической обработки заключается в удобном и лаконичном описании свойств исследуемой совокупности или явления при минимальной потере эмпирической информации. Для этого часто используется *параметрический подход*, который предполагает приближенную аппроксимацию распределения генеральной совокупности наиболее подходящим *теоретическим распределением* (нормальным, логнормальным, биномиальным, гипергеометрическим или иным). В результате формируется математическая модель $M(x, \boldsymbol{\theta})$, где x – текущее значение случайного исследуемого признака, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$ – k-мерный параметр, характеризующий заданное теоретическое распределение с учетом эмпирических данных. В общем случае вектор $\boldsymbol{\theta}$ представляет собой набор неслучайных величин (констант), значения которых не известны до эксперимента. Предполагается, что параметры $\boldsymbol{\theta}$ полностью определяют эмпирическое распределение данных в получаемых выборках.

Задача *статистического оценивания* неизвестных параметров $\boldsymbol{\theta}$ заключается в нахождении для их истинных величин наиболее точных приближенных значений: $\hat{\boldsymbol{\theta}}(x_1,...,x_n) = \{\hat{\theta}_1(x_1,...,x_n),...,\hat{\theta}_k(x_1,...,x_n)\}$ с использованием имеющихся наблюдений. Например, если предварительный анализ природы исходных данных привел нас к выводу, что их распределение может быть описано *нормальной моделью*, то для характеристики исследуемой случайной величины нам необходимо оценить только два параметра: математическое ожидание $\theta_1 = m$ и дисперсию $\theta_2 = \sigma^2$. Вся информация об этих параметрах содержится в двух выборочных статистиках – среднем арифметическом и выборочной дисперсии, которые при $n \to \infty$ сходятся по вероятности к соответствующим истинным значениям m и σ^2. Если есть основания считать, что признак имеет распределение Пуассона, то необходимо найти только один параметр λ, которым это распределение определяется.

Основной задачей оценки параметров является анализ, как будут меняться их эмпирические оценки $\hat{\theta}(\mathbf{x})$ при проецировании на всю генеральную совокупность. В силу погрешности измерений или нарушения требований независимости и случайности выборочных значений обычно мы имеем некоторое смещение $\Delta = E\{\theta - \hat{\theta}(\mathbf{x})\}$ оценки относительно истинной величины параметра θ.

Возникает вопрос о требованиях, которые следует предъявлять к статистическим оценкам, чтобы они были в определенном смысле надежными. Для этого исследуются такие свойства оценок, как *состоятельность* (т.е. $\hat{\theta}(\mathbf{x})$ стремится к θ при $n \to \infty$), *несмещенность* (т.е. $\hat{\theta}(\mathbf{x})$ совпадает с θ в среднем) и *эффективность* (т.е. $\hat{\theta}(\mathbf{x})$ обладает наименьшей степенью случайных отклонений от θ). Для измерения этих свойств разработан довольно четкий протокол параметрического оценивания, который начинается с задания формы распределения, а заканчивается выявлением связи распределения выборочных статистик и искомых параметров. Сюда обычно включаются различные

математические процедуры, такие, как метод моментов, поиск оптимума функции максимального правдоподобия, анализ цензурированных или взвешенных статистик и т.д.

Однако вопрос о том, являются ли полученные оценки параметров *наилучшими*, остается открытым. Во-первых, соображения, из которых задается вид теоретического распределения, часто оказываются субъективными. Это нередко приводит к тому, что, выполнив вычисления, мы полагаем, что найдены оценки истинных характеристик генерального распределения, тогда как на самом деле получены оценки параметров некого теоретического распределения, которое может быть «похоже» на распределение генеральной совокупности (а может быть и нет). Во-вторых, здесь можно усмотреть дефект аргументации, построенной на основе замкнутой логики: мы пользуемся выборочными характеристиками, чтобы выполнить оценивание параметров, но при этом априори задаемся предположениями о законе распределения для всей популяции, чтобы найти корректный способ расчета этих эмпирических статистик. Наконец, теоретические распределения в чистом виде в реальных условиях просто-напросто не существуют.

Имея одну конкретную выборку, можно рассчитать только один определенный набор значений оценок $\hat{\boldsymbol{\theta}}(\mathbf{x})$ параметров (например, среднее и стандартное отклонение), т.е. получить *точечные* значения статистик. При этом мы ничего не можем сказать об устойчивости модели и ошибке ее воспроизводимости: нет никакой гарантии, что при взятии повторной выборки расхождение искомых величин окажется слишком велико. Один из способов проверить корректность статистической модели состоит в том, чтобы извлекать из нашей генеральной совокупности все новые и новые повторные выборки, пересчитывать на этой основе оценки параметров и анализировать динамику изменения величины смещения Δ. Пусть, например, было взято 100 однородных независимых повторностей выборок объема n изучаемой случайной переменной. Тогда 100 значений анализируемой выборочной статистики (например, среднего) образует эмпирическое распределение, описывающее некоторую неопределенность, ассоциированную с выборочным процессом.

И здесь уместно поставить вопрос о точности и надежности найденных характеристик. При статистическом приближении, как правило, не существует гарантированной точности: нельзя указать такое $\varepsilon > 0$, для которого достоверно выполняется соотношение $\varepsilon > |\theta - \hat{\theta}(\mathbf{x})|$. Мы можем говорить лишь о вероятности, с которой выполняется это неравенство. Если эта вероятность близка к 1, то можно говорить, что статистическая погрешность в определении θ не превосходит ε.

Зададимся *доверительной вероятностью* $\gamma = 1 - 2\alpha$, близкой к 1, которой соответствуют границы области возможных значений оценок, такие, что:

$$P[\hat{\theta}(\mathbf{x}) - \varepsilon < \theta < \hat{\theta}(\mathbf{x}) + \varepsilon] = \gamma. \qquad 1.1$$

Левая $[\hat{\theta}(\mathbf{x}) - \varepsilon]$ и правая $[\hat{\theta}(\mathbf{x}) + \varepsilon]$ границы окаймляют *доверительный интервал* – статистическую меру, позволяющую с надежностью γ указать, в каких пределах может находиться случайное значение выборочной характеристики. Если анализируется несколько параметров, то доверительные границы будут окаймлять эллипс или многомерный эллипсоид. Чем больше значение γ, тем шире становится доверительная область, поэтому на практике его принято рассчитывать для нескольких доверительных вероятностей (например: 0.7, 0.9, 0.95, 0.99), чтобы получить наглядное представление о точности статистического приближения $\hat{\theta}(\mathbf{x}) \approx \theta$.

Соотношение (1.1) позволяет утверждать, что с вероятностью γ случайная доверительная область, построенная по эмпирическим данным, накроет неслучайное значение θ. На этом основании найденный доверительный интервал выборочной статистики становится интервальной оценкой соответствующего параметра θ генеральной совокупности, что позволяет перейти к более обоснованному доверительному или *интервальному* оцениванию.

Получение интервальных оценок легко осуществить при наличии большого количества повторных выборок: доверительные границы выделяются на графике эмпирического распределения $\hat{\theta}(\mathbf{x})$ для анализируемых выборок как α- и (1 - α)-квантили вариационного ряда, т.е. тогда 100(1 - 2α) из 100 реализаций результирующей статистики будет находиться внутри доверительной области. Разумеется, в реальных условиях 100 однородных и независимых повторностей получить трудно, поэтому границы доверительной области оцениваются по статистическим формулам, исходя из предположений о законе распределения случайной величины.

Существует два параметрических подхода к построению интервальных оценок (Айвазян, Мхитарян, 1998). Первый подход универсален и основан на асимптотических свойствах случайной величины $\xi(n) = (\theta - \hat{\theta})\sqrt{n}$, которая согласно утверждению ЦПТ (Центральная Предельная Теорема) распределена нормально с нулевым средним и дисперсией, зависящей от неизвестного параметра θ. Если это так, то нижняя θ_H и верхняя θ_B границы доверительного интервала оцениваемого параметра θ при доверительной вероятности $\gamma = 1 - 2\alpha$ будут равны соответственно:

$$\theta_H = \hat{\theta} - u_\alpha \sqrt{D(\hat{\theta})} \text{ и } \theta_B = \hat{\theta} + u_\alpha \sqrt{D(\hat{\theta})}, \qquad 1.2$$

где u_α - квантиль ранга α распределения N(0, 1); $D(\hat{\theta})$ – дисперсия точечной оценки.

Так как дисперсия $D(\hat{\theta})$ также обычно является неизвестной, приходится прибегать к дополнительным предположениям относительно свойств распределения уже самих данных, в частности, что они могут быть приближенно аппроксимированы тем или иным стандартным распределением.

Второй подход удается реализовать, если можно подобрать такую статистику от результатов наблюдений (x_1, x_2, …, x_n), закон распределения вероятностей которой обладает одновременно следущими свойствами: (а) не зависит от оцениваемого параметра θ; (б) описывается одним из стандартных табулируемых распределений (χ^2, F, t Стьюдента); (в) из того факта, что значения данной статистики заключены в определенных пределах с заданной вероятностью, следует, что оцениваемый параметр тоже должен лежать между некоторыми границами с той же самой вероятностью. Например, доверительные интервалы для математического ожидания m при предположении о нормальном распределении выборки из n независимых реализаций случайной величины имеют вид:

$$\bar{x} - t_\alpha s/\sqrt{n} \le m \le \bar{x} + t_\alpha s/\sqrt{n}, \qquad 1.3$$

где $\bar{x}$ и s – точечные выборочные значения среднего и стандартного отклонения; t_α – процентная точка распределения Стьюдента для $\alpha = (1 - \gamma)/2$ и (n - 1) степеней свободы.

Параметрические методы оценки доверительных интервалов во многих случаях обладают несомненной надежностью и прекрасной теоретической проработанностью. Однако на практике эти методы приходится применять и в тех случаях, когда наблюдения не вполне отвечают основным постулатам о свойствах случайной величины (независимость, распределение по Гауссу, равенство дисперсий), что придает вычислениям заведомо приближенный и даже некорректный характер. Возможные отклонения от этих предположений, характерные для экологических или экономических данных, могут серьезно повлиять на обоснованность конечных выводов: привести к смещению оценок, доверительных границ и коэффициентов связи. Часто эти нарушения бывает трудно обнаружить по ограниченному числу наблюдений и они опасны именно этой незаметностью. При явном же отличии распределения от гауссианы корректное доверительное оценивание и проверка гипотез о параметрах перерастает в сложную проблему. В таких случаях разумно вообще отказаться от стандартной нормальной модели и применять непараметрические методы, основанные на идеях Монте-Карло.

1.2. Непараметрические методы статистики и ресамплинг

Непараметрическими называют такие методы статистики, которые не зависят от какого-нибудь распределения из теоретического семейства (например, гауссового) и не используют его свойства. Они опираются лишь на предположение, что случайная величина X независима и тождественно распределена. В результате своей простоты и универсальности непараметрические модели приобрели широкую область применения и составили эффективную конкуренцию традиционным параметрическим методам.

Первоначально непараметрические методы предназначались для проверки статистических гипотез. После работ А.Н. Колмогорова и открытия Ф. Уилкоксоном (1945 г.) ранговых критериев, используемых для анализа различий между одномерными выборками и степени взаимосвязи переменных, оформилось целое научное направление, основанное на использовании ранговых процедур. Другой раздел непараметрической статистики – непараметрическое оценивание характеристик неизвестных генеральных совокупностей – позволяет осуществлять построение частотных гистограмм и анализировать эмпирическую функцию плотности распределения с оценкой таких ее характеристик, как квантили, моменты, мода, энтропия, информация по Фишеру и т. д.

Попробуем прокомментировать разницу в концептуальной основе параметрических и непараметрических тестов (хотя различие между ними не вполне ясно, и вряд ли станет полностью ясным после наших рассуждений). Чтобы сопоставить эти два подхода, рассмотрим предварительно весьма прозаичный пример, связанный с подбрасыванием монеты. Мы не знаем, действительно ли кто-то увлекался этим почтенным занятием, но анализ соотношения вероятностей “аверс-реверс” очень популярен в статистике, поскольку аналогичен многим нашим практическим ситуациям.

Итак, предположим, что мы имеем дело со старыми римскими монетами, у которой на передней стороне нанесено больше серебра, чем на обратной, т.е. у них смещен центр тяжести. Мы хотим оценить вероятность того, что в результате 10 подбрасываний у нас выпадет 10 “решек”, и, если она высока, то тестируемая монета является подлинной (пример предложен Д. Ховелом).

Выполняя параметрический тест, мы формулируем нулевую гипотезу, что выпадение орла и решки являются равновероятными, и задаемся вопросом «Какова вероятность выпадания 10 решек из 10 подбрасываний, если нулевая гипотеза H_0: $p = 0.50$ является истиной?». Подбросим последовательно 10 заведомо подлинных монет, зафиксируем результат и выполним некоторые простые вычисления, основанные на биномиальном распределении и описанные во многих учебниках по статистике. Еще раз обратим внимание, что мы априори постулировали некоторое теоретическое распределение, оценили его *параметр* (π) и задали тем самым статистическую модель, основанную на этом параметре.

При непараметрическом подходе мы задаемся более лаконичным вопросом «Если монета подлинна, то как часто мы получаем 10 решек в результате 10 бросков?». Этот вопрос не использует слово "параметр" и не нуждается в априорных предположениях, что вероятность решки на любом броске в случае нулевой гипотезы равна $p = 0.50$. Также нет необходимости проводить какие-либо аналогии с биномиальным распределением. Мы просто берем 100 подлинных монет и будем их одновременно общей большой кучей 10 раз подбрасывать вверх (как это сделать – чисто техническая проблема, не связанная со статистикой). И мы можем легко подсчитать, какое количество монет из 100 легло лицевой стороной вверх все 10 раз, и принять эту вероятность как норму при оценки “подлинности или фальшивости” валюты.

В общем случае непараметрические подходы, связанные с построением эмпической функции распределения статистических характеристик изучаемой случайной величины, возможны только при наличии повторностей наблюдений, в нашем примере – 10 серий по 100 старинных римских монет. Однако в экологии (как и в экономике, медицине и многих других отраслях) можно выполнить срез данных только в

определенном месте и в определенный момент времени, а если отбирать вторую, третью пробы и т.д., то это будут уже данные из другого места или же взятые в другой момент времени. Поэтому возникает вопрос: как, имея лишь одну единственную повторность, оценить значение необходимого нам показателя и получить меру точности этой оценки?

В том случае, когда нет возможности получить истинные повторности наблюдений**,** разработаны методы, которые формируют большое количество так называемых "*псевдовыборок*", и на их основе можно получить необходимые характеристики искомого параметра: оценки математического ожидания, дисперсии, доверительного интервала. Методы "численного ресамплинга" или, как их иногда называют в русскоязычной литературе, "методы генерации повторных выборок" объединяют четыре разных подхода, отличающихся по алгоритму, но близких по сути: рандомизация, или перестановочный тест (permutation), бутстреп (bootstrap), метод "складного ножа" (jackknife) и кросс-проверка (cross-validation). Эти алгоритмы, моделирующие эмпирическое распределение выборочных характеристик, являются современной альтернативой параметрическим методам и бурно развиваются два последних десятилетия.

То, что параметрические и непараметрические методы не являются антагонистами, а могут вполне гармонично дополнять друг друга, особенно ярко проявляется на примере бутстрепа. Обычно его реализация не требуют никакой априорной информации о законе распределения изучаемой случайной величины, однако использование различных алгоритмов параметризации при генерации псевдовыборок может явиться решающим приемом стабилизации результатов.

Процедуры ресамплинга выполняют обработку различных фрагментов исходного массива эмпирических данных, как бы поворачивая их разными гранями и сопоставляя полученные таким образом результаты (Эфрон, 1988). Вопрос о полной корректности такого приема остается открытым, но если признать его законным, то асимптотические достоинства ресамплинга удается доказать вполне строго. Генерируемые псевдовыборки, вообще говоря, не являются независимыми, однако при увеличении их числа влияние зависимости может ослабевать и с ресамплированными значениями статистик можно обращаться как с независимыми случайными величинами, считая их вполне состоятельными оценками параметров (Орлов, 2007).

Отметим встречающиеся в литературе различия в русскоязычной транслитерации применяемых терминов. Так "ресамплинг" (resampling) часто используют в более близкой к английскому языку фонетической транскрипции: "ресэмплинг" или "ресемплинг". Иногда процедуру численного ресамплинга называют "перевыборкой" (впрочем, этот термин иногда используют и как синоним "псевдовыборки" или "псевдореплики"). Ю.П. Адлером, редактором книги Эфрона (1988), был введен термин "бутстреп", тогда как в последнее время стало принято использовать "бутстрап" и "бутстрапинг" (Анатольев, 2007). Рандомизационный тест часто называют перестановочным или пермутационным (permutation), хотя некоторые авторы (Zieffler et al., 2011, p. 133) видят между ними концептуальные различия в природе порождения данных: были ли они случайно выбраны (randomly sampled) или случайно назначены (randomly assigned). Кросс-проверка (а также "перекрестная проверка" или "кросс-валидация") была впервые теоретически обоснована под названием "скользящий контроль" (Вапник, Червоненкис, 1974). Надеемся, что читатели со снисхождением отнесутся к авторам в части выбираемых ими вариантов терминологии.

1.3. Складной нож и бутстреп – механизмы генерации случайных псевдовыборок

Идеи численного ресамплинга не являются принципиально новыми в статистике и относятся по крайней мере к 1935 году, но практическое применение этих методик было связано с вынужденым ожиданием пока не появятся достаточно быстрые компьютеры.

Один из первых алгоритмов, предложенный М. Кенуем в 1949 г., заключался в том, чтобы последовательно и многократно исключать из имеющейся выборки,

насчитывающей n элементов, по одному ее члену и обрабатывать вариационный ряд из оставшихся (n - 1) элементов. Среднее значение, дисперсия или медиана будут при этом “блуждать” и тогда можно проанализировать информацию о каждом акте смещения, построить распределение выборочной оценки искомого параметра и уточнить его свойства. Дж. Тьюки активно усовершенствовал этот метод, назвав его "*jackknife*" (складной нож), и использовал для оценки дисперсии изучаемой совокупности и проверки нулевой гипотезы о том, что распределение некоторой статистики симметрично относительно заданной точки. «Понятие "складной нож" относится к универсальному методу, призванному заменить частные методики, которые не всегда пригодны, подобно бойскаутскому ножу, годящемуся на все случаи жизни» (Мостеллер, Тьюки, 1982, с. 143).

Предположим мы имеем выборку **X** из 6 элементов {3.12; 0; 1.57; 19.67; 0.22; 2.2} со значениями среднего арифметического $\bar{x}$ = 4.46 и стандартного отклонения s = 7.54. Традиционные параметрические методы позволяют нам оценить точность оценки $\bar{x}$ или ошибку среднего:

$$s_m = s/\sqrt{n} = \left[\frac{1}{n(n-1)}\sum_{i=1}^{n}(x_i - \bar{x})^2\right]^{1/2} = 3.08. \qquad 1.4$$

Но ту же ошибку среднего мы можем вычислить, исходя из иных соображений. Отбрасывая в исходной выборке по одному члену, сформируем шесть псевдовыборок по 5 элементов в каждой. Из этих данных, в которых поочередно была исключена i-я точка, получим ряд выборочных средних $\tilde{x}_{(i)}$: {4.73; 5.36; 5.04; 1.42; 5.31; 4.92}. В нашем случае после полного перебора всех возможных вариантов среднее арифметическое из “обрезанных” средних $\bar{x}_{(\bullet)}$ будет в точности равно исходному среднему $\bar{x}$ = 4.46, что можно доказать алгебраически.

Использование метода складного ножа не имеет смысла, когда необходимо найти возможное смещение меры положения, но оно потенциально весьма полезно во многих других случаях оценки параметров распределения. Поскольку каждая из сформированных псеводвыборок также является подмножеством из той же генеральной совокупности, то оценка для ошибки среднего, вычисленная методом складного ножа, будет равна:

$$\hat{\sigma}_{JACK} = \left[\frac{n-1}{n}\sum_{i=1}^{n}(\tilde{x}_{(i)} - \bar{x}_{(\bullet)})^2\right]^{1/2} = 3.017. \qquad 1.5$$

Смещение параметра, полученное jackknife-процедурой по сравнению с вычислениями по обычным формулам, оказалось незначительным и это может выглядеть как «…чудовищное разбазаривание вычислительных усилий. С содержательной точки зрения процедуру складного ножа можно трактовать лишь как попытку подменить теоретический анализ вычислительной мощью. Но за счет этого нам удается освободиться от ограничений традиционной параметрической теории с ее повышенным доверием к малому набору стандартных моделей, допускающих теоретическое решение» (Эфрон, 1988, с. 50).

Другой аргумент связан с тем, что формула (1.4) не допускает очевидных обобщений на какие-либо иные оценки параметров, кроме среднего $\bar{x}$, такие как медиана, стандартное отклонение, асимметрия, эксцесс и проч. Преимущество формулы (1.5) в том, что она легко обобщается на любую статистику $\theta_n = f_n(x_1, x_2, \ldots, x_n)$. Для этого достаточно лишь сделать подстановку $\tilde{\theta}_{(i)}$ и $\bar{\theta}_{(\bullet)}$ вместо $\tilde{x}_{(i)}$ и $\bar{x}_{(\bullet)}$ соответственно. Например, для шести псевдовыборок с исключенным i-м членом стандартные отклонения $\tilde{s}_{(i)}$ будут принимать значения { 8.4; 8.07; 8.28; 1.32; 8.11; 8.34} при среднем $\bar{\theta}_{(\bullet)}$ = 7.09. Подставив эти значения в формулу (1.5), получим оценку складного ножа для ошибки стандартного отклонения $\hat{\sigma}_{JACK}$ = 5.66.

Популярность метода "складного ножа" с его недостаточно интенсивным вычислительным подходом при анализе выборочных оценок параметров существенно снизилась в ходе развития идей бутстрепа, когда появилась возможность гибкой настройки и использование алгоритмов самоорганизации. Вместе с тем, jackknife-методы нашли в экологии широкое применение для прогнозирования числа "невидимых" редких видов и экстраполяции видового богатства сообществ (см. раздел 5.1).

Идеи складного ножа получили дальнейшее развитие на общий случай эмпирического оценивания параметров любых моделей регрессии или распознавания, построенных по прецедентам, в рамках процедуры кросс-проверки (cross-validation), подробно представленной в разделе 3.4.

Бутстреп-процедура (или *bootstrap***)** была предложена (Efron, 1979б) как некоторое обобщение алгоритма "складного ножа", чтобы не уменьшать каждый раз число элементов по сравнению с исходной совокупностью. По одной из версий слово "bootstrap" означает кожаную полоску в виде петли, прикрепляемую к заднику походного ботинка для облегчения его натягивания на ногу. Благодаря этому термину появилась английская поговорка 30-х годов: «Lift oneself by the bootstrap», которую можно трактовать как «Пробить себе дорогу благодаря собственным усилиям» (или подобно барону Мюнхгаузену вытянуть себя из болота за шнурки от ботинок).

Основная идея бутстрепа по Б. Эфрону (1988) состоит в том, чтобы методом статистических испытаний Монте-Карло многократно извлекать повторные выборки из эмпирического распределения. А именно: берется конечная совокупность из n членов исходной выборки $x_1, x_2, \ldots, x_{n-1}, x_n$, откуда на каждом шаге из n последовательных итераций с помощью датчика случайных чисел, равномерно распределенных на интервале [1, n], "вытягивается" произвольный элемент x_k, который снова "возвращается" в исходную выборку (т.е. может быть извлечен повторно). Например, при $n = 6$ одна из таких комбинаций имеет вид $x_4, x_2, x_2, x_1, x_4, x_5$, т.е. одни элементы могут повторяться два или более раз, тогда как другие элементы отсутствовать.

Таким способом можно сформировать любое, сколь угодно большое число бутстреп-выборок (обычно 5000-10000). Как и в случае "складного ножа", в результате легкой модификации частотного распределения реализаций исходных данных можно ожидать, что каждая следующая генерируемая псевдовыборка будет возвращать значение параметра, немного отличающееся от вычисленного для первоначальной совокупности. На основе разброса значений анализируемого показателя, полученного в процессе имитации, можно построить, например, доверительные интервалы оцениваемого параметра. Тем самым бутстреп представляет собой более экономный способ статистического исследования, использующий всю вычислительную мощь компьютера, но позволяющий обойтись без дополнительных натурных измерений.

Бутстреп, как и иные методы генерации повторных выборок, полезны, когда статистические выводы нельзя получить с использованием теоретических предположений (например, какие-либо предположения сделать трудно из-за недостаточного объема выборок). Они незаменимы, чтобы оценить степень устойчивости или неопределенности оценок относительно наблюдаемых данных. Наконец, они могут использоваться, чтобы просто проверить полноценность стандартных приближений параметрическими моделями и улучшить их, если выяснится, что они дают неадекватные результаты.

В зависимости от имеющейся информации относительно статистической модели генеральной совокупности различают непараметрический и параметрический бутстреп. В общем виде *непараметрическая* бутстреп-процедура выглядит следующим образом:

Шаг 1: Получение большого количества повторностей – случайных наборов данных из изучаемой совокупности. В качестве исходных данных берется, как правило, только одна случайная выборка, полученная эмпирическим путем. Вместо того, чтобы делать новые повторности эксперимента, на основе одной имеющейся выборки генерируется множество псевдовыборок того же размера, состоящих из случайных комбинаций исходного набора

элементов. При этом используется алгоритм "случайного выбора с возвращением" (random sampling with replacement), т.е. извлеченное число снова помещается в "перемешиваемую колоду" прежде чем вытягивается следующее наблюдение. В результате некоторые члены в каждой отдельной псевдовыборке могут повторяться два или более раз, тогда как другие – отсутствовать. Отметим, что если бы мы осуществляли выбор без возвращения (random sampling without replacement), то все время получали бы исходное множество чисел, хотя и представленное каждый раз в различном порядке.

На рис. 1.1 представлены три псевдовыборки из исходного набора шести наблюдений. Практически, разумеется, генерируют сотни или тысячи псевдовыборок, а не только три. Естественными ограничениями здесь являются желаемые затраты компьютерного времени и отчасти размер эмпирической совокупности *n*. Однако если при использовании складного ножа мы можем сгенерировать лишь *n* псевдовыборок, то в случае бутстрепа число возможных вариантов носит комбинаторный характер.

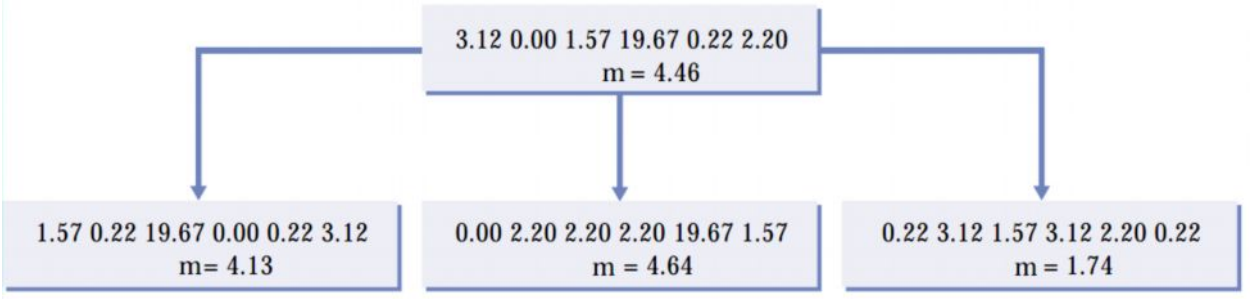


Рис. 1.1. Идея бутстрепа. Верхний блок включает эмпирическую выборку из 6 значений. Три нижних блока являются псевдовыборками, случайно составленными из элементов исходной выборки. Рассчитана статистическая характеристика – выборочное среднее *m* – как для исходной совокупности, так и для каждой псевдовыборки.

Шаг 2: Построение бутстреп-распределения оцениваемой величины. Для каждой псевдоповторности, полученной на шаге 1, рассчитывается значение анализируемой характеристики – среднего, медианы, стандартного отклонения и др. Имея это множество данных, легко построить гистограмму (или сглаженный график плотности частотного распределения) значений тестируемого показателя, отражающую закономерности его вариации, что дает возможность оценить доверительные интервалы и другие полезные выборочные характеристики анализируемой величины.

Если простой непараметрический бутстреп выполняет перевыборку с учетом *равной вероятности* появления каждого элемента, то *стратифицированный* бутстреп учитывает соотношение частот между относительно гомогенными группами (стратами), на которые может быть разделены выборочные объекты. Речь идет о генерации весов f^*_{bsk}, регулирующих частоту появления в *b*-й бутстреп-реплике *k*-го выборочного значения, относящегося к *s*-й страте, что может быть выполнено различными методами. Так Х. Саиго (Saigo, 2010) сравнил для случая трехуровневой стратификации исходной выборки достоинства и недостатки четырех возможных алгоритмов бутстрепа: а) зеркального соответствия (Mirror-Match Bootstrap); б) бутстрепа без возвращения (Without-Replacement Bootstrap); в) бутстрепа Бернулли и г) бутстрепа с изменением масштаба (Rescaling Bootstrap).

Поскольку непараметрический бутстреп в общем виде можно представить как извлечение псевдовыборки из дискретного эмпирического распределения выборки наблюдений, то это используется для поиска возможности улучшить характеристики этого распределения различными методами сглаживания. Например, выберем величину *h* и будем корректировать формируемую бутстреп-выборку путем добавления к каждому *i*-му ее члену случайного приращения $x^*_i \pm h\varepsilon_i$, где $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ – реализации независимой,

нормально распределенной N(0, 1) случайной величины. Такой подход называется бутстрепом со сглаживанием (Smooth bootstrap).

Параметрический бутстреп использует предположение, что исходные выборочные данные представляют собой случайные реализации вероятностного процесса, определяемого некоторым теоретическим распределением. Выбор конкретной модели является самостоятельной важной задачей, решению которой посвящено большое число литературных источников. Предположим, что эта проблема уже решена, и нам необходимо найти наиболее правдоподобное значение произвольной статистики *T*, которая является некоторой функцией от *X*. Тогда процедура параметрического бутстрепа состоит из следующих шагов:

<u>*Шаг 1*</u>. По выборочным данным $\{x_1, x_2, \ldots, , x_n\}$ осуществляется построение вероятностной модели и оцениваются ее параметры $\hat{\boldsymbol{\theta}}$ (например, $\bar{x}$ и σ в случае нормального распределения).

<u>*Шаг 2*</u>. Случайным образом из подобранного распределения с параметрами $\hat{\boldsymbol{\theta}}$ генерируются *n* элементов $\{x^*_1, x^*_2, \ldots, , x^*_n\}$ и бутстреп-повторность, полученная такой имитацией, используется для расчета значения статистики $t^* = T(\mathbf{x}^*)$.

<u>*Шаг 3*</u>. Шаг 2 выполняется *B* раз и формируется бутстреп-распределение анализируемой статистики $\{t^{*1}, t^{*2}, \ldots, t^{*j}, \ldots, t^{*B}\}$.

Байесовский бутстреп вместо того, чтобы имитировать выборочное распределение статистической величины, оценивающей параметр, моделирует апостериорное распределение самого параметра. Поскольку при этом методы формирования выводов основаны на нескольких специфических модельных предположениях, то полученные результаты весьма чувствительны к этим предположениям.

Существуют мнения (Manly, 2007), что сам бутстреп (как и рассматриваемая в следующей главе рандомизация) является частным случаем испытаний *Монте-Карло* **(**см. первые работы Бюффона в 1777 г.), в которых специфический стохастический процесс осуществляет равновероятные перестановки данных между уровнями воздействия. При обоих подходах псевдовыборки, сгенерированные в ходе имитации, используются для моделирования распределений, оценки доверительных интервалов или проверки нулевой гипотезы. Особенно близко эта связь прослеживается при использовании байесовских методов – см. марковские цепи Монте-Карло в разделе 7.7.

1.4. Оценка среднего и доверительных интервалов бутстреп-методом.

В общем случае оценка параметров по выборочным данным осуществляется исходя из природы предполагаемого вероятностного процесса порождения данных, т.е. задаются предположениями, что выборка случайно извлечена из генеральной совокупности, описываемой некоторой теоретической функцией распределения $F(x)$.

При моделировании непрерывными распределениями важнейшей задачей статистической обработки является оценка математического ожидания реализаций случайной величины $\boldsymbol{X}$: $EX = \mu(F) = \int x dF(x)$ и ее дисперсии $\mathrm{D}X = \sigma^2(F) = \mathrm{E}(X - \mathrm{E}X)^2$.

Оценить значения μ можно на основе данных наблюдений с помощью выборочного среднего $m(\hat{F})$. И здесь типичны следующие проблемы: как велико смещение и вариация оценки *m*? Каковы реальные доверительные интервалы μ? Совместимы наши предположения о законе распределения $F(x)$ с данными наблюдений?

Рассмотрим предварительно два следующих примера.

Предположим, что требуется оценить среднее видовое богатство макрозообентоса в одной пробе из р. Байтуган [пример 2 из приложения 1, в дальнейшем обозначаемый как П2]. Исходная выборка включает данные по $n = 60$ выполненным пробам, в которых число обнаруженных видов варьирует от 2 до 30 при среднем их числе $\bar{x} = 11.2$. Расчет по известным теоретическим формулам дает следующие значения описательных статистик:

стандартное отклонение $s = 5.69$, ошибка среднего $s_m = 0.735$ и двухсторонние доверительные интервалы среднего с надежностью $\gamma = 95\%$ и $t_\gamma = 2$

$$9.73 = 11.2 - 2 \cdot 0.735 \geq \bar{x} \geq 11.2 + 2 \cdot 0.735 = 12.67,$$

т.е. с вероятностью 95% можно предположить, что в одной пробе нами будет найдено от 9.73 до 12.67 видов донных организмов.

Для выполнения процедур ресаплинга воспользуемся простой и удобной программой Resampling Procedures 1.3, разработанной и свободно распространяемой Д. Ховелом (см. Предисловие). Выполним $b = 5000$ итераций бутстрепа и для каждой сгенерированной j-й псевдовыборки из $\{\boldsymbol{x}^{*1}, \boldsymbol{x}^{*2}, \ldots, \boldsymbol{x}^{*j}, \ldots, \boldsymbol{x}^{*b}\}$ размерностью n вычислим частные величины среднего $\bar{x}^{*j}$ и стандартного отклонения s^{*j}. Тогда улучшенные общие значения среднего $\bar{x}_{boot}$ и стандартной ошибки бутстрепа se_{boot} могут быть вычислены по обычным формулам усреднения частных значений:

$$\bar{x}_{boot} = \frac{1}{b}\sum_{j=1}^{b}\bar{x}^{*j} = 11.2; \qquad se_{boot} = \left[\frac{1}{b-1}\sum_{j=1}^{b}(\bar{x}^{*j} - \bar{x}_{boot})^2\right]^{1/2} = 0.723. \qquad 1.6$$

На основе бутстрепированных данных, полученных в результате имитации, легко построить гистограмму частотного распределения $\bar{x}^{*j}$ (рис. 1.2) и найти граничные значения точечного богатства видов при 95% доверительной вероятности, не используя при этом предположений о нормальном характере распределения.

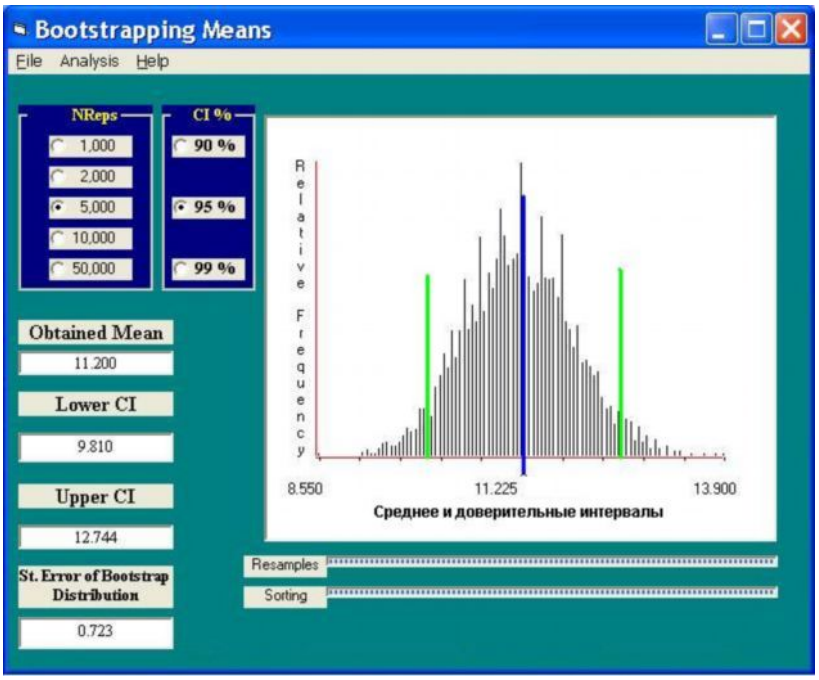


Рис. 1.2. Оценка параметров распределения числа видов в пробах бентоса с использованием модуля Resampling Procedures 1.3 (Howell, 2001).

Здесь и далее: NReps – число генерируемых псевдовыборок; CI – доверительная вероятность; Obtained Mean –среднее для исходной выборки (линия синего цвета на гистограмме – среднее, скорректированное бутстреп-методом); Lower и Upper CI – нижняя и верхняя границы доверительного интервала соответственно (зеленые линии на гистограмме); St.Error of Bootstrap Distribution – стандартное отклонение бутстреп-распределения.

Очевидно, что в этом примере выборочные статистики, рассчитанные по классическим формулам, почти совпадают с полученными бутстрепом. И возникает вопрос: если можно просто подставить исходные данные в простую формулу, зачем было выполнять непрактичную работу по обработке 5000 случайных рядов?

Может также показаться подозрительным, что генерация повторных выборок создает новые данные «как бы из ничего». Но в бутстрепе псевдовыборки и не

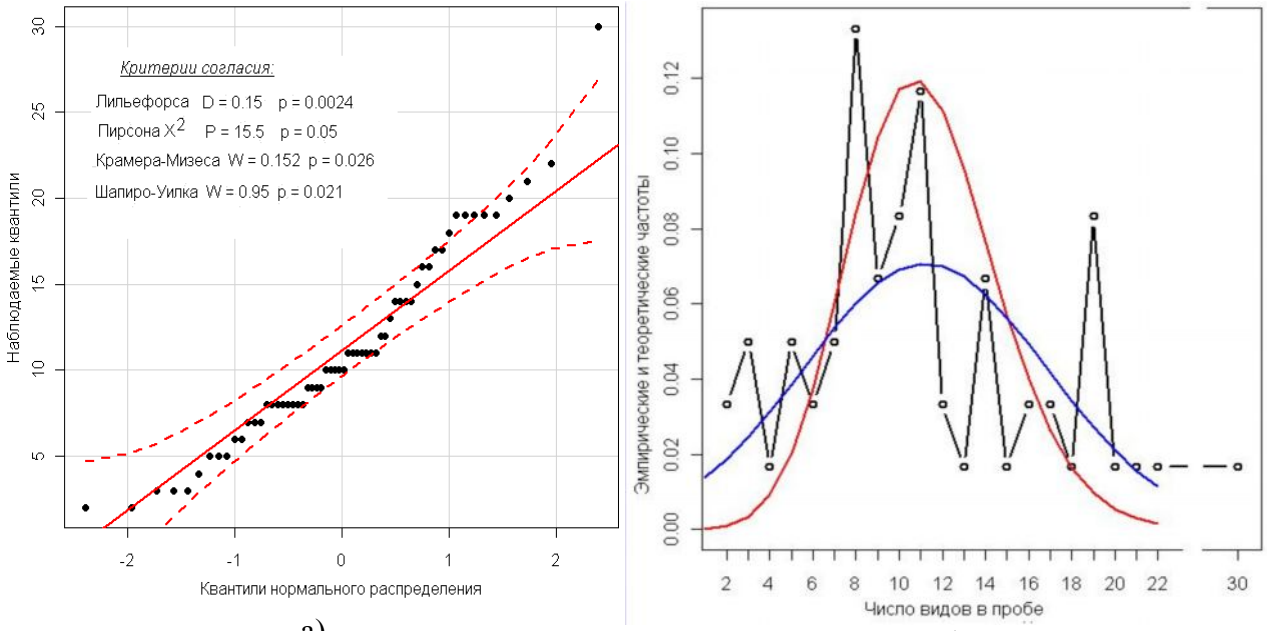
воспринимаются как новые данные – распределение средних $\bar{x}^{*j}$ используется лишь для того, чтобы оценить, как выборочное среднее одной эмпирической выборки размером 60 изменилось бы в результате легких случайных флуктуаций. Это совершенно законно позволяет детально проанализировать дрейф среднего и меру его вариации, опираясь на основные статистические формулы (1.6), и не полагаться на возможно неверное предположение, что ошибка среднего равна $s_m = s/\sqrt{n}$.

Отметим здесь важную деталь: стандартная ошибка бутстрепа se_{boot} для среднего является одновременно стандартным отклонением бутстреп-распределения этого же статистического параметра. Это правило легко обобщается на любую иную выборочную характеристику $\hat{\theta} = \theta(x_1, x_2, \ldots, x_n)$: имея разброс имитированных значений θ^{*j} для b выборок бутстрепа, можно по единой общей формуле рассчитать ее стандартную ошибку:

$$se_{boot} = \left[\frac{1}{b-1}\sum_{j=1}^{b}[\theta^{*j} - \theta^{*}(\cdot)]^2\right]^{1/2}, \text{ где } \theta^{*}(\cdot) = \frac{1}{b}\sum_{j=1}^{b}\theta^{*j} \quad \text{(Efron, Tibshirani, 1993).} \qquad 1.7$$

По сути рассмотренного примера следует сделать два важных замечания.

Во-первых, для анализируемого ряда гипотеза о нормальности распределения отвергается (см. рис. 1.3а) и использование статистики Стьюдента при оценке доверительных интервалов является здесь не вполне неправомочной. Однако, имея в своем распоряжении выборки более 100 самых разнообразных показателей эколого-биологического профиля (см. Приложение 1), мы не смогли найти ни одной (sic!), где бы принималась гипотеза о нормальном законе распределения. Повторим опять: нормального распределения в чистом виде в реальных условиях просто-напросто не существует. Поэтому в центре внимания исследователя должен стоять вопрос не о том, нарушены ли предположения статистического анализа, а *как далеко можно зайти в их нарушении*.



а) б)

Рис. 1.3. График квантиль-квантильного нормального распределения числа видов в пробе (а) и аппроксимация эмпирических частот этого события распределением Пуассона (кривая красного цвета, $\lambda = 11.2$) и нормальным распределением (синяя кривая $\mu = 11.2$, $\sigma = 5.69$) (б)

Практическое экспериментирование приводит к размыванию жесткости предпосылок и откровениям типа: «Нормальность превратилась не более чем в частный случай» (Дрейпер, Смит, 1986, с. 12); либо вовсе: «Нормальный закон, как закон ошибок неверен» (Тутубалин и др., 1999; с. 8). Было показано (Орлов, 2007), что при больших

объемах выборок требование нормальности ослабевает (нужный эффект обеспечивается центральной предельной теоремой ЦПТ) и оценка доверительных интервалов с помощью критерия Стьюдента дает правильные результаты, независимо от того, выполняются ли предпосылки нормальности или нет.

Использование бутстрепа делает еще один шаг к анализу возможностей применения статистической формулы (1.3) при нарушенных предположениях о нормальности. Очевидно, что если это условие не подтверждается, но выборка является достаточно большой, то бутстреп-распределение выборочной характеристики $\bar{x}$ будет чаще всего подчиняться нормальному закону (см. рис. 1.2) со стандартным отклонением $s/\sqrt{n}$, т.к. тут работает ЦПТ. И это открывает широкие возможности по анализу доверительных интервалов. Если не касаться непростого вопроса приоритетов, то логично утверждать: если значения доверительных интервалов, полученных параметрическим методом и с использованием ресамплинга, в приемлемой степени совпадают, то это является веским аргументом, что этой интервальной оценке параметра можно доверять.

Во-вторых, можно увидеть проблему в том, что выборку, состоящую из дискретных значений количества видов, мы аппроксимируем непрерывным нормальным распределением. Представим, что наблюдаемая случайная величина X является последовательностью из чисел независимых событий, произошедших за время исследований с постоянной интенсивностью $\lambda = 11.2$ (λ – параметр распределения Пуассона). Какое из двух рассматриваемых теоретических распределений – нормальное $f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{(x-\mu)^2/2\sigma^2}$ или пуассоновское $f(x) = \frac{\lambda^x}{x!} e^{-\lambda}$ – лучше приближено к распределению эмпирических частот (см. рис. 1.3б) и какую параметрическую модель следует принять? Строгое решение этой проблемы не является очевидным, и мы приведем только значения критериев Колмогорова-Смирнова для проверки согласия эмпирических $\hat{F}_n(x)$ и теоретических частот: $D_n = 0.227$ ($p = 0.62$) при сравнении с нормальным распределением и $D_n = 0.4$ ($p = 0.05$) при сравнении с распределением Пуассона.

Рассмотрим второй пример, связанный с оценкой параметров по выборке массы тела 213 особей ящерицы прыткой *Lacerta agilis* (П5). Характер распределения этого вариационного ряда также весьма далек от нормального закона, в чем легко убедиться, построив квантиль-квантильный график (рис. 1.4). Рассчитаем все интересующие нас выборочные статистики: среднее $\bar{x} = 13.65$, стандартное отклонение $s = 6.06$, коэффициент вариации $CV = 6.06/13.65 = 0.444$, ошибку среднего $s_m = 0.415$ и, несмотря на отсутствие необходимых предпосылок нормальности, двухсторонние доверительные интервалы среднего с надежностью $\gamma = 95\%$ (п. 1 табл. 1.1):

$$12.83 = 13.65 - 2.02 \cdot 0.415 \geq \bar{x} \geq 13.65 + 2.02 \cdot 0.415 = 14.46.$$

Представим набор возможных методов оценки интервальных значений искомых параметров с использованием ресамплинга на примере среднего $\bar{x}$ и коэффициента вариации[2] $CV = s/\bar{x} = \frac{1}{\bar{x}}\left[\sum_{i=1}^{n}(x_i - \bar{x})^2/(n-1)\right]^{1/2}$ для выборки массы тела ящериц. Отметим предварительно, что ошибка коэффициента вариации, расчитанная по приближенной формуле $s_{CV} = CV/n^{1/2} = 0.03$, а двухсторонние доверительные интервалы с надежностью $\gamma = 0.95$ равны $CI_{0.95} = CV \pm t_{0.025}\, s_{CV} = 0.444 \pm 2.02 \cdot 0.03 = 0.384 \div 0.504$.

[2] Как показал Е.Л. Воробейчик (1993), коэффициент вариации лежит в основе математических формул большинства индексов разнообразия, дистанции, ширины/перекрывания экологической ниши, обсуждаемых нами далее

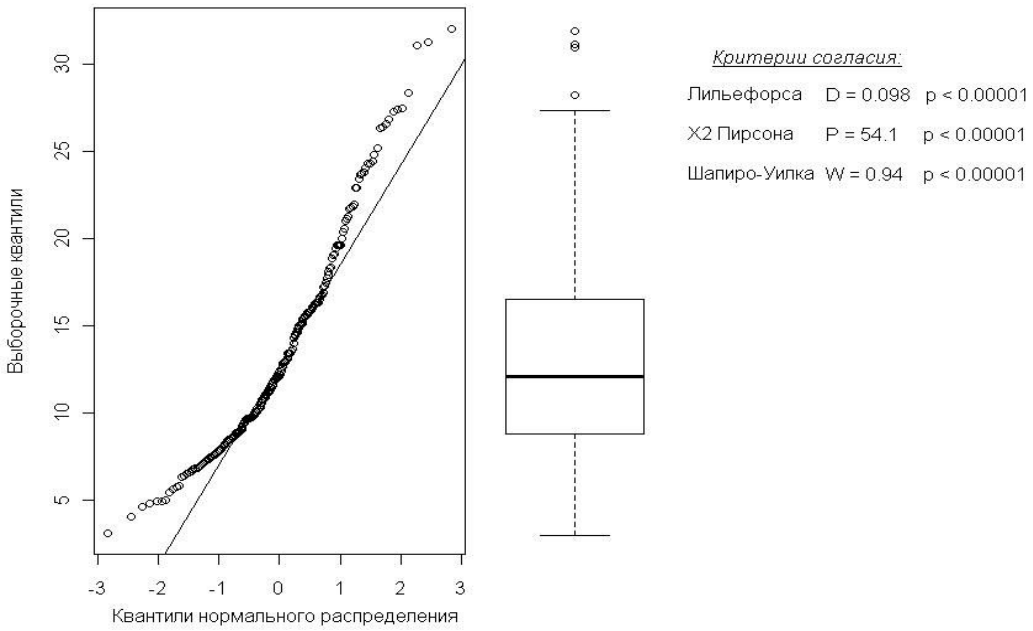


Рис. 1.4. График квантиль-квантильного нормального распределения выборки массы тела ящерицы прыткой

При выполнении бутстрепа на основе имеющихся у нас эмпирических данных сгенерируем, используя алгоритм замены с возвращением, 5000 псевдовыборок и вычислим для каждой из них оценки математического ожидания и коэффициента вариации. На основе полученных вариационных рядов $\{\bar{x}^{*1}, \bar{x}^{*2}, \ldots, \bar{x}^{*5000}\}$ и $\{CV^{*1}, CV^{*2}, \ldots, CV^{*5000}\}$ построим графики функции распределения $\hat{F}$ (рис. 1.5) и найдем средние значения этих бутстреп-статистик:

- среднего $\bar{x}_{boot}$ = 13.65 (смещение оценок равно 0.0058);
- коэффициента вариации CV^*_{boot} = 0.4426 с абсолютным смещением 0.0014 (0.3%).

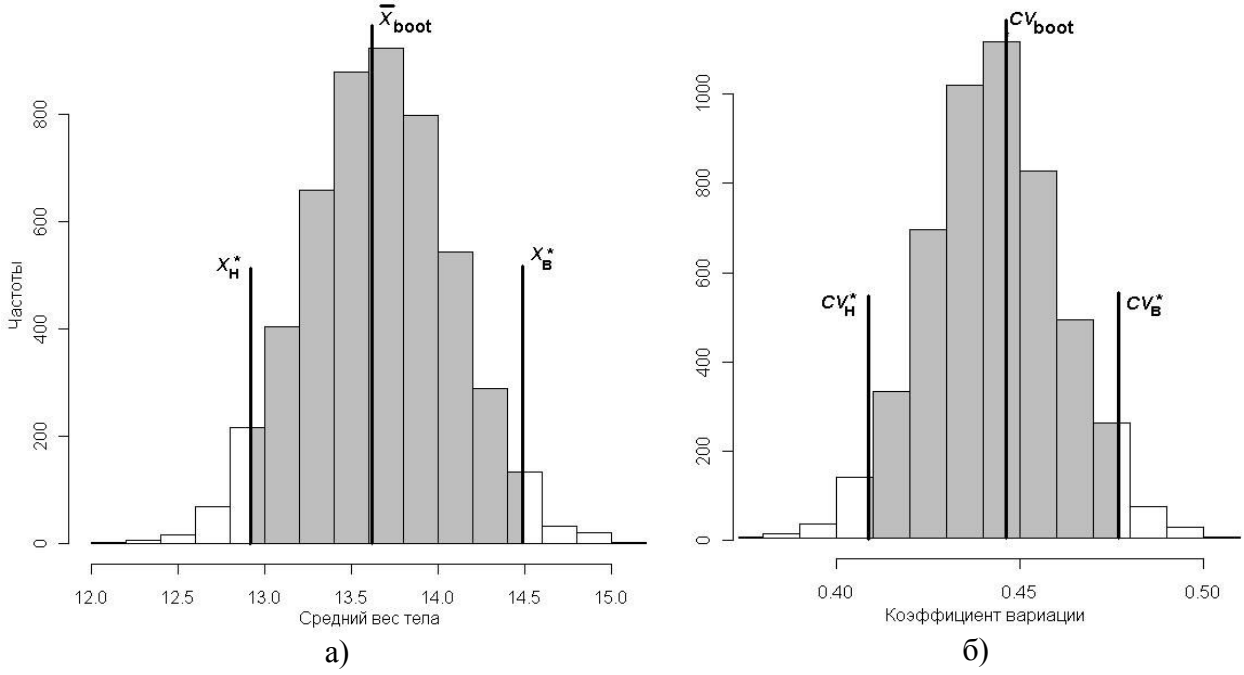


Рис. 1.5. Гистограммы частотного распределения средних (а) и коэффициентов вариации (б), полученные бутстрепированием выборки массы тела ящериц

Существуют различные теоретические соображения (Everitt, Howell, 2005, pp. 169-176; Davison, Hinkley 2006), чтобы получить интервальные оценки параметров с использованием бутстрепа. Основными условиями реализации этих методов являются симметричность и унимодальность распределения исходной выборки, а также алгоритм бутстрепирования, обеспечивающий генерацию из этого распределения случайных и независимых повторностей.

Наиболее простым и естественным является *метод процентилей* (п. 3 табл. 1.1). Если выбран уровень γ, соответствующий доверительной вероятности (1- 2α), то нижнюю X^*_H и верхнюю X^*_B границы интервалов, удовлетворяющие γ, можно приблизительно найти по следующим соотношениям:

$$\gamma = I(X^*_H \le x^{*j} \le X^*_B)/B; \quad \alpha = I(x^{*j} \le X^*_H)/B = I(x^{*j} \ge X^*_B)/B; \qquad 1.8$$

где B – объем бутстреп-повторений; $I(X^*_H \le x^{*j} \le X^*_B)/B$ – количество оценок среднего $\bar{x}^{*j}$ из общего числа B бутстреп-повторений, принявших значения в диапазоне между X^*_H и X^*_B (или за пределами этих границ в формуле для α). Иными словами, в методе процентилей в качестве границ доверительного интервала (bootstrap percentile interval) среднего будут квантили $[q^*_\alpha = X^*_H,\ q^*_{(1-\alpha)} = X^*_B]$ бутстреп-распределения, т.е. при γ = 0.95 – 125-е (или 0.025·5000) и 4875-е (или 0.975·5000) значения упорядоченного вариационного ряда статистических величин $\bar{x}^{*j}$, полученных в ходе имитации. Очевидно, что при снижении надежности статистического вывода до 90%, ими будут 250-е и 4750-е значения и доверительный интервал сужается.

Таблица 1.1. Оценки доверительных интервалов и средней массы тела ящериц и коэффициента вариации, полученные различными методами

Использованный метод оценки	Надежность, %	Среднее	Коэффициент вариации *CV*
1. По исходной выборке с использованием *t*-критерия (предположения не соблюдаются)	95	12.83 ÷ 14.46	0.384 ÷ 0.504
2. Метод "складного ножа" (jackknife)	95	12.83 ÷ 14.46	0.408 ÷ 0.481
3. Метод процентилей	95	12.84 ÷ 14.48	0.407 ÷ 0.479
	90	12.97 ÷ 14.34	0.412 ÷ 0.473
4. Основные интервалы (basic *CI*)	95	12.81 ÷ 14.45	0.407 ÷ 0.478
5. Бутстреп с использованием *t*-критерия	95	12.86 ÷ 14.46	0.409 ÷ 0.482
6. Интервалы стьюдентизированного типа	95	12.91 ÷ 14.56	0.410 ÷ 0.483
7. С коррекцией смещения BCa	95	12.90 ÷ 14.49	0.411 ÷ 0.484

Метод процентилей выглядит вполне разумным, однако имеются два серьезных возражения. Первое, вытекающее из теоретических соображений, говорит о том, что здесь мы фактически находим не доверительные интервалы для искомого параметра, а толерантные интервалы выборочных реализаций бутстрепируемой статистики. Иными словами, найдя значения процентилей $\bar{x}^{*(2.5\%)}$ и $\bar{x}^{*(97.5\%)}$, мы можем сделать заключение, что среднее из любой комбинации наших эмпирических данных с вероятностью 95% укладывается в эти границы, но это не вполне соответствует искомой оценке уверенности относительно значения параметра μ.

Другое возражение связано с возможной отчетливой асимметрией распределения бутстреп-статистик для эмпирических выборок. Например, на рис. 1.5 расстояния от среднего до доверительных границ составляют $l_н = \bar{x}_{boot} - X^*_H = 0.6744$ и $l_в = X^*_B - \bar{x}_{boot} = 0.6734$. К.Луннеборг (Lunneborg, 2000) анализирует эту ситуацию и показывает, что, если использовать среднее $\bar{x}$ как анализируемую статистику, то при заданной доверительной вероятности оценка параметра будет статистически значима в пределах от $(\bar{x} - l_в)$ до $(\bar{x} + l_н)$, тогда как с использованием метода процентилей эти границы вычисляются с

точностью до наоборот, т.е. от ($\bar{x}$ - $l_н$) до ($\bar{x}$ + $l_в$). Тогда *основные доверительные интервалы* бутстрепа (basic bootstrap interval) для среднего будут вычисляться по формуле

$$[2\bar{x}_{boot} - q^*_{(1-\alpha)}; 2\bar{x}_{boot} - q^*_{\alpha}] \quad (1.9)$$

– см. п. 4 табл. 1.1. Отметим, что для симметричных распределений $l_н = l_в$ эта проблема не имеет практического значения и границы процентильных интервалов совпадают с основными.

Метод оценки доверительных интервалов, использующий *t*-распределение Стьюдента, основан на теоретическом утверждении ЦПТ, что при $n \to \infty$ искомый параметр имеет нормальное распределение, а его бутстреп-оценка доставляет минимальное смещение относительно истинного значения. Тогда, например, доверительную область *C*, включающую математическое ожидание μ с надежностью $\gamma = (1 - 2\alpha)$, можно рассчитать по обычной статистической формуле (1.3):

$$\bar{x}_{boot} \pm t_\alpha se_{boot}, \quad (1.10)$$

где среднее $\bar{x}_{boot}$ и стандартная ошибка бутстрепа se_{boot} рассчитаны по выражению (1.5), а t_α является критическим значением α-го квантиля распределения Стьюдента $t(\alpha, n - 1)$, ограничивающим доверительную область *C* – см. п. 5 табл. 1.1.

Заметим все же, что Госсет первоначально получил *t*-распределение при условии, что анализируемая случайная величина распределена нормально, поэтому истинные доверительные границы могут иметь некоторый сдвиг, пропорциональный тому, насколько конкретная эмпирическая выборка отклоняется от этого предположения. Метод стьюдентизированных доверительных интервалов (п. 6 табл. 1.1) ставит своей целью скомпенсировать этот сдвиг, отказавшись от предположения о нормальности распределения $\hat{F}$ бутстреп-оценок, и скорректировать критические значения t_α. Вспомним, что выполняя *B* итераций бутстрепа, мы вычисляли для каждой *i*-й сгенерированной псевдовыборки значения среднего $\bar{x}^{*i}$ и стандартного отклонения s^{*i}. На основе этих статистических данных мы можем вычислить бутстрепированные значения $t^*_i = (\bar{x}^{*i} - \bar{X}) / S_{\bar{X}}{}^*$ и восстановить функцию распределения t^*, не использующую предположения о нормальности. Нам теперь остается только найти по гистограмме характерные значения t^* для 97.5 % и 2.5 %-х вероятностей и заменить ими критическую величину t_α в традиционной формуле. Мы получаем доверительные границы

$$X^*_H = 2\bar{x}_{boot} - t^*_{0.975} se_{boot} \text{ и } X^*_B = 2\bar{x}_{boot} - t^*_{0.025} se_{boot}.$$

Заметим, что мы поменяли местами 2.5 и 97.5-ые процентили t^* по той же причине, по которой это сделано для формулы основных интервалов (1.9).

Надо сказать, что выше приведены не самые лучшие варианты решения по компенсации сдвига. Эфрон 20 лет посвятил этой проблеме и разработал процедуру BC_a (bias correction and acceleration) коррекции доверительных границ, которая учитывает различные выбросы, дрейф стандартной ошибки среднего и другие факторы – п. 7 табл. 1.1. Процедура слишком громоздка, чтобы обсуждать ее здесь, но она подробно описана в одном из самых полных учебных пособий по бутстрепу (Efron, Tibshirani, 1993).

Отметим (не делая категоричных выводов) близость оценок доверительных интервалов в табл. 1.1, полученных бутстрепом и с использованием *t*-критерия при нарушенных предположениях о нормальности.

 [3]

К разделу 1.4:

```
#  Варианты бутстреп-анализа и оценка доверительных интервалов
#  Загрузка данных (числа видов в пробе) из текстового файла
T <- read.delim("Species.txt") ; x <- T$S ; n = length(x)
summary(x)  #  вывод основных выборочных статистик
```

[3] Некоторые предварительные замечания по использованию скриптов можно найти в Приложении 2

```
# Вывод в окно графика квантиль-квантилей нормального и выборочного распределений
library(car) ; qqPlot(x, dist= "norm",xlab="Квантили нормального распределения",
                                        ylab="Наблюдаемые квантили", main="", pch=19)
#  Проверка нормальности по критериям  Лильефорса, χ² Пирсона, Шапиро-Уилка и Крамера-Мизеса
library(nortest) ;  lillie.test(x) ; pearson.test(x) ;  shapiro.test(x) ; cvm.test(x)
# Оценка параметров нормального распределения и распределения Пуассона
p1 = mean(x) ; p2 = sqrt(var(x)*(n-1)/n)
# Создание векторов эмпирических и теоретических частот
pr_obs <- as.vector(table(x)/length(x)) ; nr <- length(pr_obs)
pr_norm <- dnorm(1:nr, p1, p2) ; pr_pois <- dpois(1:nr, p1)
plot(table(x)/length(x),type="b") #  Отрисовка графика на рис. 1.3б
lines(1:nr, pr_pois , col="red", lwd=2) ; lines(1:nr, pr_norm, col="blue", lwd=2)
# Проверка согласия по критерию Колмогорова-Смирнова
ks.test(pr_obs, pr_norm) ; ks.test(pr_obs, pr_pois)
# ---------------------------------------------------------------------------
# Загрузка выборочных данных по массе тела ящерицы Zootoca из текстового файла
Z <- read.delim("Zootoca.txt") ; x = Z$bm ; summary(x); n <- length(x); mean(x) ; sd(x)
#  Оценку доверительных интервалов проводим на примере коэффициента вариации
#   Для оценки значений среднего необходимо изменить функции CV()  и f.CV()
CV <- function(x) sqrt(var(x))/mean(x)  # определение функции для коэффициента вариации
CV(x)     # расчет коэффициента вариации для исходной выборки
CV(x)+ qt(0.025,df=n-1)*CV(x)/sqrt(n) ; CV(x)+ qt(0.025,df=n-1)*CV(x)/sqrt(n)
# С помощью функции sample(…, replace=T) создаем случайные выборки с «возвращением»
# Коэффициенты вариации для 5000 бутстреп-выборок помещаем в массив bt
bt <-numeric(5000); for (i in 1:5000) bt[i] <- CV(sample(x, replace=T))
# Визуализация бутстреп-распределения и оценка смещения
hist(bt); mean(bt) - CV(x)
#  оценка доверительных интервалов методом процентилей
quantile(bt, prob=c(0.025,0.975))   # с 95% доверительной вероятностью
quantile(bt, prob=c(0.05,0.95))       #  с 90% доверительной вероятностью
#  оценка доверительных интервалов с использованием t-критерия – функция gt(α, df)
mean(bt) - qt(0.975,length(x)-1)*sqrt(var(bt))
mean(bt) + qt(0.975,length(x)-1)*sqrt(var(bt))
#  оценка доверительных интервалов методом basic CI
2*mean(bt) - quantile(bt,0.025)  ;  2*mean(bt) - quantile(bt,0.975)
library(boot)  #   Загрузка специализированного пакета boot
#  Основная функция бутстрепинга - boot(), имеющая  следующий формат:
#  Bootobject <- boot(data= , statistic=, R=, ...), где data – вектор или таблица с данными
#  statistic – функция, возвращающая  k бутстепируемых статистик; k ≥ 1
#  R - число  бутстреп-реплик
# определение функции для коэффициента вариации
f.CV <- function(y,id) {sqrt(var(y[id]))/mean(y[id])}
bootres <- boot(x, f.CV, 5000) # накопление результатов в Bootobject
print(bootres) ; plot(bootres) # вывод текстовых и графических результатов
# вывод комплекта из различных версий доверительных интервалов
boot.ci(bootres, conf = 0.95, type = c("norm", "basic", "perc", "bca"))
library(bootstrap)   #   Загрузка специализированного пакета bootstrap
# нахождение 95% стьюдентизированных доверительных интервалов
bootres2=boott(x, f.CV, VS=TRUE, v.nboott=5000, perc=c(.025,.975)) ; bootres2[1]
# нахождение 95% непараметрических ABC-доверительных интервалов
thetaCV <- function(p, x)  {xm <- sum(p*x)/sum(p)
       den <- sum((p*x - sum(p*x)/sum(p))^2)/(length(x)-1)
       return(sqrt(den)/xm)}
abcnon(x, thetaCV)
# нахождение 95% доверительных интервалов методом складного ножа
jack <- numeric(length(x)-1) ; pseudo <- numeric(length(x))
for (i in 1:length(x))
     { for (j in 1:length(x))
       {if(j < i) jack[j] <- x[j] else if(j > i) jack[j-1] <- x[j]}
     pseudo[i] <- length(x)*CV(x) -(length(x)-1)*CV(jack)}
mean(pseudo) - qt(0.975,length(x)-1)*sqrt(var(pseudo)/length(x))
mean(pseudo) + qt(0.975,length(x)-1)*sqrt(var(pseudo)/length(x))
```

1.5. Подбор параметров распределений и примеры параметрического бутстрепа

С помощью бутстрепа можно делать то, что не всегда под силу обычным параметрическим методам. Например, для асимметричных выборок часто предлагают использовать медиану в качестве оценки меры положения случайной величины вместо традиционного математического ожидания.

Если распределение данных близко к нормальному, то стандартную ошибку медианы можно приближенно оценить по формуле $\sigma_{med} = 1.253\ \sigma / \sqrt{n}$, т.е. считается, что она на 25 % больше, чем ошибка среднего σ. При существенных отклонениях от нормальности ошибку медианы или моды обычными способами рассчитать трудно из-за отсутствия повторностей выборок. Бутстреп-метод дает возможность сгенерировать из исходной выборки, предположим, 5000 псевдомедиан, что позволяет легко рассчитать и стандартную ошибку медианы, и ее доверительные интервалы.

Рассмотрим выборку популяционной плотности (экз/м^2) личинок *Chironomus salinarius*[4] в 43 пробах из различных рек Самарской обл. [пример П2], где этот вид был обнаружен (в скобках приведены частоты повторяющихся значений):

5 (2); 8; 10 (3); 19; 20 (3); 30; 40; 42; 50 (6); 65 (2); 80 (3); 100 (2); 133; 200; 250; 300; 430; 440; 480; 800; 880; 2400; 3020; 3360; 5200; 6200; 7000; 9000; 19000.

Характер распределения этого вариационного ряда весьма далек от нормального закона, что дает основания предполагать, что медиана является одной из наиболее интерпретируемых оценок меры положения. Найденное выборочное значение медианы Me = 80, что, кстати, значительно меньше среднего арифметического ($\bar{x}$ = 1432).

Выполним непараметрический бутстреп представленной выборки и оценим доверительные интервалы медианы численности *Chironomus salinarius* (рис. 1.6).

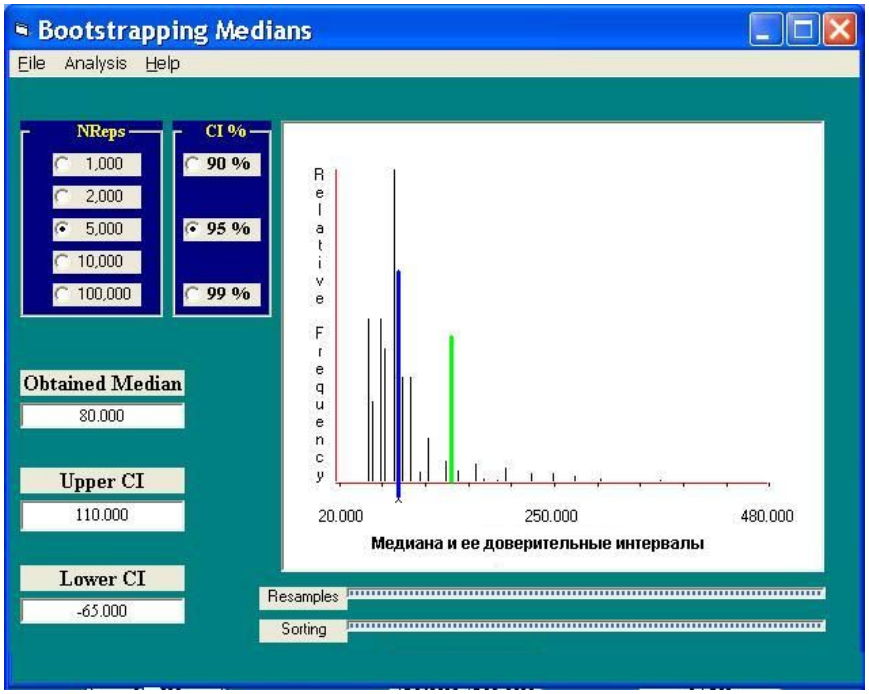


Рис. 1.6. Оценка доверительных интервалов медианы численностей *Chironomus salinarius* бутстреп-методом

[4] Червевидные красные личинки (мотыль) комаров-звонцов вида *Chironomus salinarius*; как и другие личинки хирономид являются превосходным кормом для рыб

В общем случае, если мы имеем результаты наблюдений, измеренные в шкалах численности объектов, то количество зарегистрированных уникальных значений, как правило, невелико. В нашем случае, как бы много не проводилось итераций бутстрепа, мы можем получить только 43 псевдозначения медиан. Оценка стандартного отклонения по такому частотному распределению связана с существенной статистической ошибкой, поэтому нижняя граница 95%-го доверительного интервала, рассчитанного некоторыми методами, основанными на se_{boot}, может сместиться в отрицательную область (см. рис. 1.6). Для сравнения в представленном примере доверительные интервалы, рассчитанные методом процентилей равны 50 ÷200, а методом BCa – 50 ÷133.

Описанная проблема характерна для многих приложений, где измеряемый показатель представлен в баллах: классы качества водоемов, оценки знаний учащихся (Цейтлин, 2012) и т.д. Ее решение может быть осуществлено с использованием одного из методов имитации Монте-Карло. В частности, генерация любой по объему последовательности случайной величины из заданного распределения легко осуществляется на основе алгоритма обратной трансформации (Inverse Transform Method – см. например, Rubinstein, Kroese, 2003).

Пусть X – непрерывная случайная величина имеет кумулятивную функцию распределения $F_X(x)$, т.е. для каждого x значение $F(x)$ равно вероятности того, что любая произвольная величина ξ из этого распределения будет меньше x: $F(x) = P(\xi < x)$. Так как $F_X(x)$ является неубывающей функцией, то обратная функция $F_X^{-1}(y)$ может быть определена для любого значения y между 0 и 1. При этом $F_X^{-1}(y)$ будет равняться тому значению x, для которого $F(x) = y$. В частности, если множество U равномерно распределено по интервалу (0, 1), чтобы получить значение x случайной величины X, задаются дискретным значением u из множества U, вычисляют величину $F_X^{-1}(u)$ и принимают его равным x.

Эмпирическая функция распределения (ЭФР), полученная сглаживанием частот численности *Chironomus salinarius* в 43 пробах, представлена на рис. 1.7 (для удобства визуализации число особей представлено в логарифмической шкале). Из графика видно, что, например, вероятности $u = 0.6$ соответствует количество особей $x = 100$ (т.е. в 60% проб плотность популяции оказалось меньше, чем 100 экз/м^2).

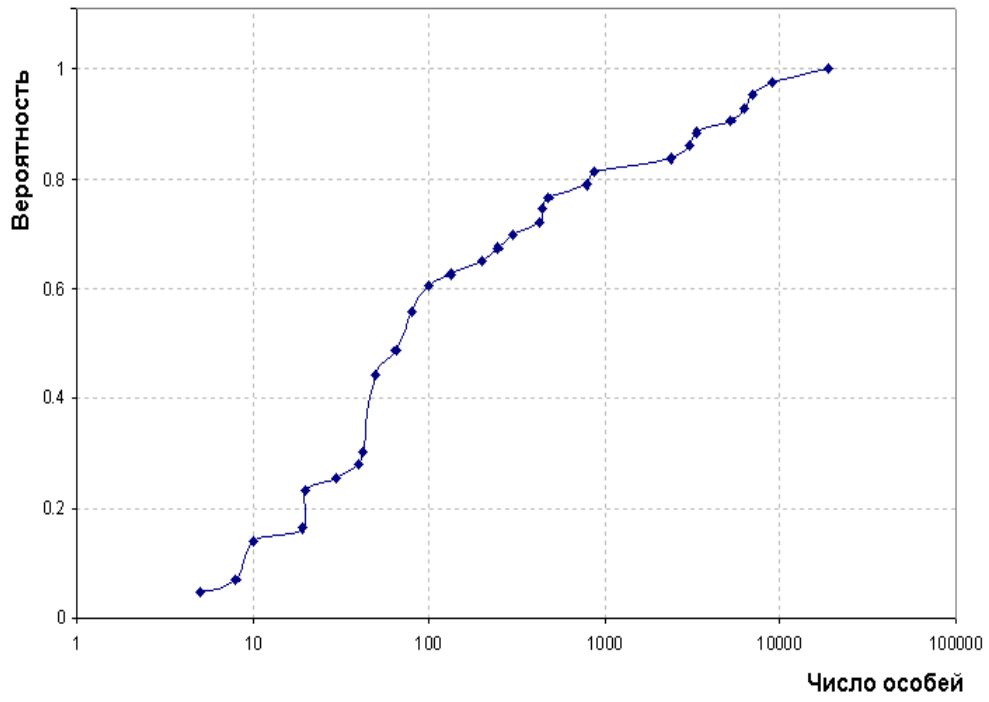


Рис. 1.7. График эмпирической функции распределения численностей экземпляров *Chironomus salinarius*

Поскольку вид теоретической функции распределения $F(x)$, как правило, неизвестен, есть, как минимум, два пути получить аналитическое выражение для ЭФР на рис. 1.7: а) кусочно-полиномиальная интерполяция выборочной ЭФР, сглаживание сплайнами и проч., б) подбор наиболее подходящего теоретического распределения.

Для уточнения доверительных интервалов медианы используем первый способ. С использованием генератора случайных чисел на интервале (0, 1) простой линейной интерполяцией получим имитируемую последовательность из 200 значений популяционной плотности в виде следующего вариационного ряда (приведены минимаксные значения и квартили):
0.028 (u = 0.00027), …, 19.9 (u = 0.228), …, 74.9 (u = 0.53), …, 456 (u = 0.753), …, 18621 (u = 0.999).
Бутстреп-процедура с использованием выборки из псевдочисленностей, полученных методом обратной трансформации, сформировала более реалистичные, чем на рис. 1.6, границы доверительных интервалов: 52.3 ÷ 98.3 (метод процентилей); 49.9 ÷ 97.2 (с использованием t-критерия) и 47.3 ÷ 93.4 (BCa).

Перейдем теперь к *параметрическому бутстрепу*. В общем случае оценка параметров по выборочным данным осуществляется следующим способом:

- задаются предположениями, что выборка случайно извлечена из генеральной совокупности, описываемой некоторой теоретической функцией распределения $F(x)$;
- по выборочным данным оцениваются параметры $\hat{\theta}$ предполагаемого распределения методом максимального правдоподобия или с использованием теоретических формул (в статистической среде R эту работу выполняет, например, функция fitdistr(…) из пакета MASS);
- с использованием критериев согласия проверяется гипотеза об отсутствии статистически значимых различий между эмпирической функцией распределения (ЭФР) $\hat{F}$ и функцией распределения $F(x)$ с параметрами $\hat{\theta}$;
- если нулевая гипотеза отклонена, то уточняют предположения о законе распределения $F(x)$, выполняют преобразования данных (например, трансформация Бокса-Кокса) или решают проблемы расщепления смеси распределений.

Как отмечалось выше, бутстреп может быть реализован в рамках заданной параметрической модели, что дает возможность эффективно оценивать статистические характеристики ее параметров. При этом псевдо-выборки формируются не путем рекомбинирования элементов исходного фиксированного ряда, а извлекаются из некоторого теоретического распределения.

Принято считать, что численность организмов X в отобранных пробах распределена по закону Пуассона: $P(X = x) = \frac{\lambda^x}{x!} e^{-\lambda}$, где λ – параметр интенсивности процесса, выборочной оценкой которого является средняя частота событий $\bar{x}$.

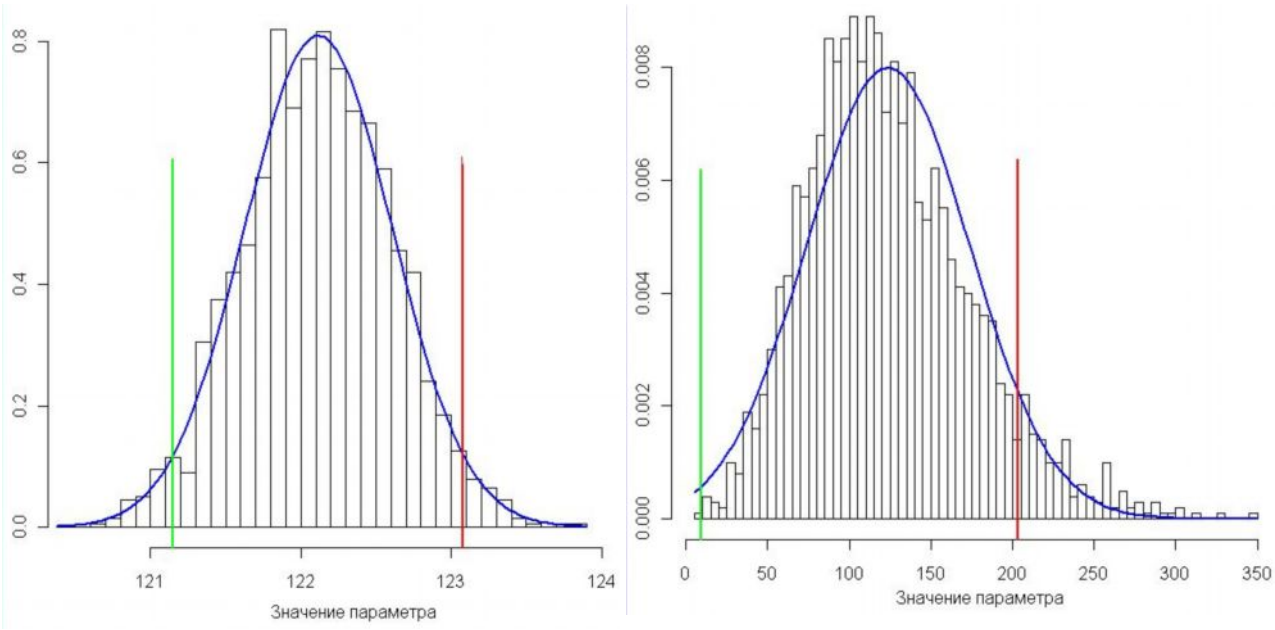
Оценка доверительных интервалов λ с использованием параметрического бутстрепа в данном случае сводится к следующему:

- рассчитывается оценка параметра λ_{obs} для выборочных данных (из соображений общности для этого обычно используется метод моментов MM или максимального правдоподобия MLE);
- из распределения Пуассона с параметром λ_{obs} извлекается случайная выборка размером n, по которой уточняется и запоминается оценка параметра λ_{sim};
- предыдущий шаг повторяется B раз, строится вариационный ряд статистик λ_{sim} и на его основе любым описанным выше методом оцениваются доверительные интервалы.

Если ограничиться эмпирическим распределением численности экземпляров *Chironomus salinarius* только в тех гидробиологических пробах, в которых этот вид удалось обнаружить, то интенсивность появления особей, рассчитанная параметрическим бутстрепом при B = 2000, составляет λ_{boot} = 1400 экз/м^2 с доверительным интервалом,

найденным методом процентилей, от 1389 до 1412 экз/м^2. Практически к тем же результатам приводит и использование обычного непараметрического бутстрепа.

Однако, выполнив все вышеприведенные расчеты, мы игнорировали 450 проб, в которых особи вида *Chironomus salinarius* вообще не встречались. Если добавить к представленной выборке из 43 численностей еще 450 значений "нулевого хвоста", то ситуация качественным образом меняется (см. рис. 1.8). Во-первых, уменьшается значение параметра λ_{obs}, который равен выборочному среднему $\bar{x}$ = 122 экз/м^2 . Во-вторых, непараметрический бутстреп, ориентированный на исходную выборку, в ходе итераций активно использует элементы "нулевого хвоста", тогда как параметрический бутстреп продолжает извлекать выборки из распределения Пуассона в довольно узком диапазоне относительно λ_{obs}. Этим объясняется большая разница в оценках доверительных интервалов, полученных этими двумя подходами.



а) б)

Рис. 1.8. Гистограммы частотного распределения и доверительные интервалы оценок параметра λ распределения Пуассона по выборке численности *Chironomus salinarius* с учетом "нулевого хвоста", полученные параметрическим (а) и непараметрическим (б) бутстрепом

Одновременно заметим, что представленный пример имеет в значительной мере учебный характер. В настоящее время появляется большое количество работ, описывающих теорию и практику построения обобщенных моделей регрессии Пуассона с "нулевым хвостом" (ZIP – Zero-Inflated Generalized Poisson Regression Model), которые являются концептуальной основой оценки вероятности встречаемости особей редких видов (Liu et al., 2011).

В большинстве экологических приложений поиск базовой вероятностной модели для выборочных данных – сложный и неоднозначный процесс. Рассмотрим подбор теоретического распределения для выборки массы тела ящерицы прыткой *Lacerta agilis* (П5), статистические характеристики которой анализировались выше (см. табл. 1.1). Эффективным способом проверки согласия распределений является построение квантиль-квантильных QQ-графиков (рис. 1.4): если квантили нормального распределения и эмпирические квантили пропорциональны между собой и выборочные точки строго выстраиваются на "теоретической" прямой, то аппроксимацию можно считать удачной. Для той же цели можно использовать и ZZ-графики стандартизованных z-значений.

Однако насколько статистически значим разброс экспериментальных точек относительно теоретической прямой QQ? Сгенерируем большое число *B* случайных

выборок из нормального распределения с параметрами, оцененными по выборочным данным. Мы будем иметь целый пучок прямых с немного отличающимися коэффициентами угла наклона. Если провести перпендикулярно через этот пучок секущие плоскости и соединить кривой крайние точки, то мы получим "коридор" из двух *огибающих* (point-wise envelope), внутри которого будет располагаться любая из сгенерированных прямых – см. рис. 1.9. Можно провести также семейство огибающих с различным уровнем доверительной вероятности, т.е. окаймляющих, например, 90% бутстреп-прямых (подробности см. в Davison, Hinkley, 1997, раздел 4.2.4).

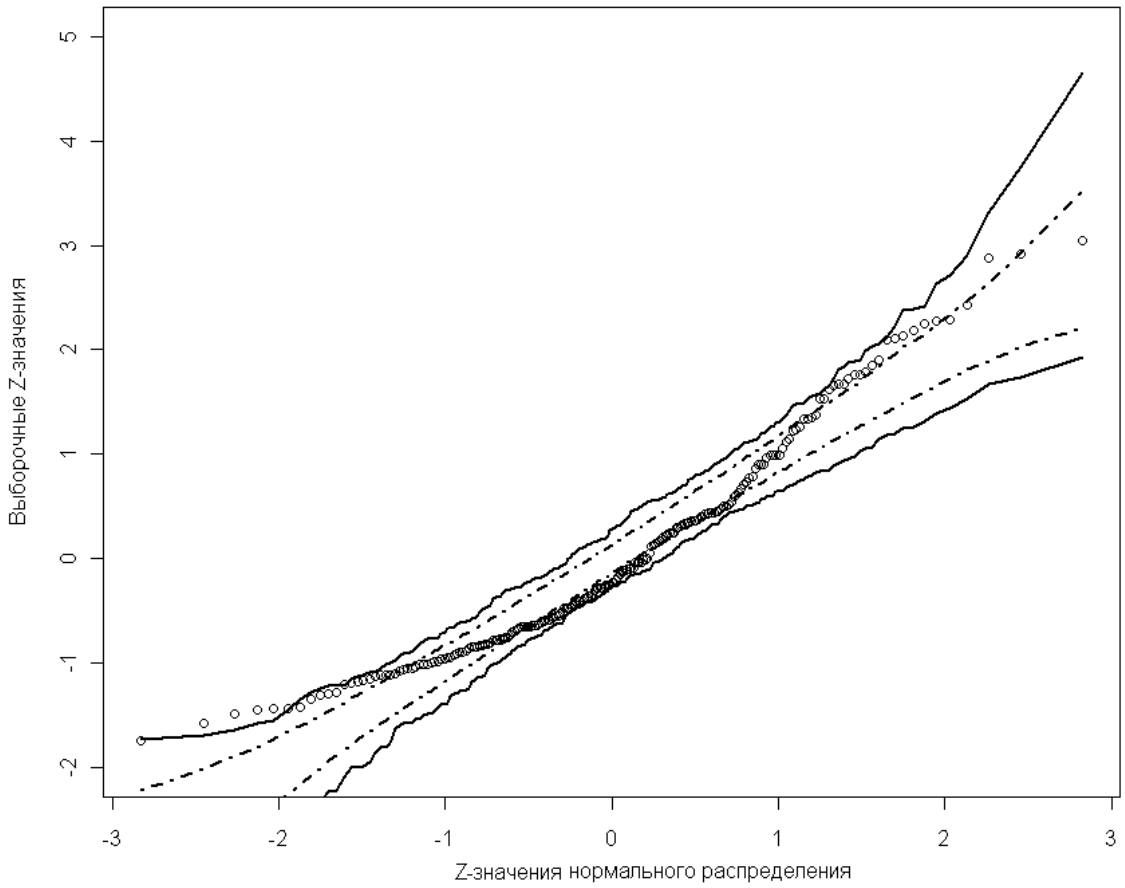
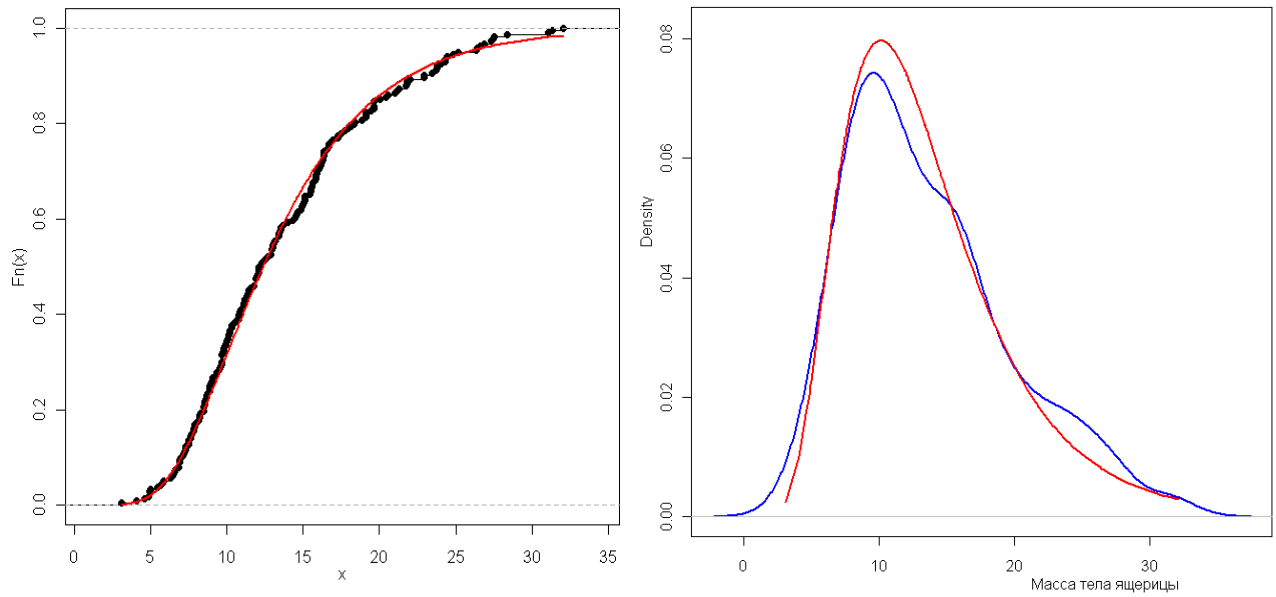


Рис. 1.9. График зависимости z-значений для нормального распределения и по наблюдаемым данным массы тела ящерицы прыткой; показаны доверительные огибающие, полученные после 2000 итераций бутстрепа (100% – сплошная линия, 90% – штрих-пунктир)

Поскольку аппроксимация нормальным распределением оказалась неудовлетворительной, рассмотрим варианты использования двух других базовых распределений, широко используемых в демографических исследованиях, – логнормального и Вейбулла. Ниже приведены оценки параметров этих распределений, их ошибки, найденные методом максимального правдоподобия, а также величины критерия согласия Колмогорова-Смирнова D и соответствующие ему p-значения:

Вид распределения	Параметры и их ошибки		D	p
Нормальное	$m = 13.64 \pm 0.41$;	$\sigma = 6.04 \pm 0.29$	0.099	0.032
Вейбулла	$\alpha = 15.44 \pm 0.46$;	$1/\lambda = 2.41 \pm 0.12$	0.0716	0.226
Лог-нормальное	$m_\eta = 2.516 \pm 0.031$;	$\sigma_\eta = 0.446 \pm 0.021$	0.0429	0.827

Уточнить характер отличий эмпирического и подбираемого теоретического распределений удобно с использованием совместного графика функций плотности распределения вероятностей ЭФПР и ТФПР – см. рис. 1.10. Вид ЭФПР может быть представлен гистограммой или функцией ядерного сглаживания (см. далее раздел 7.1).



а) б)

Рис. 1.10. Аппроксимация выборочных данных массы тела ящериц логнормальным распределением: эмпирическая и теоретическая кумулятивная функции распределения вероятностей (а); теоретическая функция плотности вероятности (красная) и ядерная функция (синяя), сглаживающая эмпирическое распределение (б)

Для логнормального распределения описанной выше бутстреп-процедурой также может быть построен квантиль-квантильный график с огибающими, ограничивающими принятые доверительные интервалы.

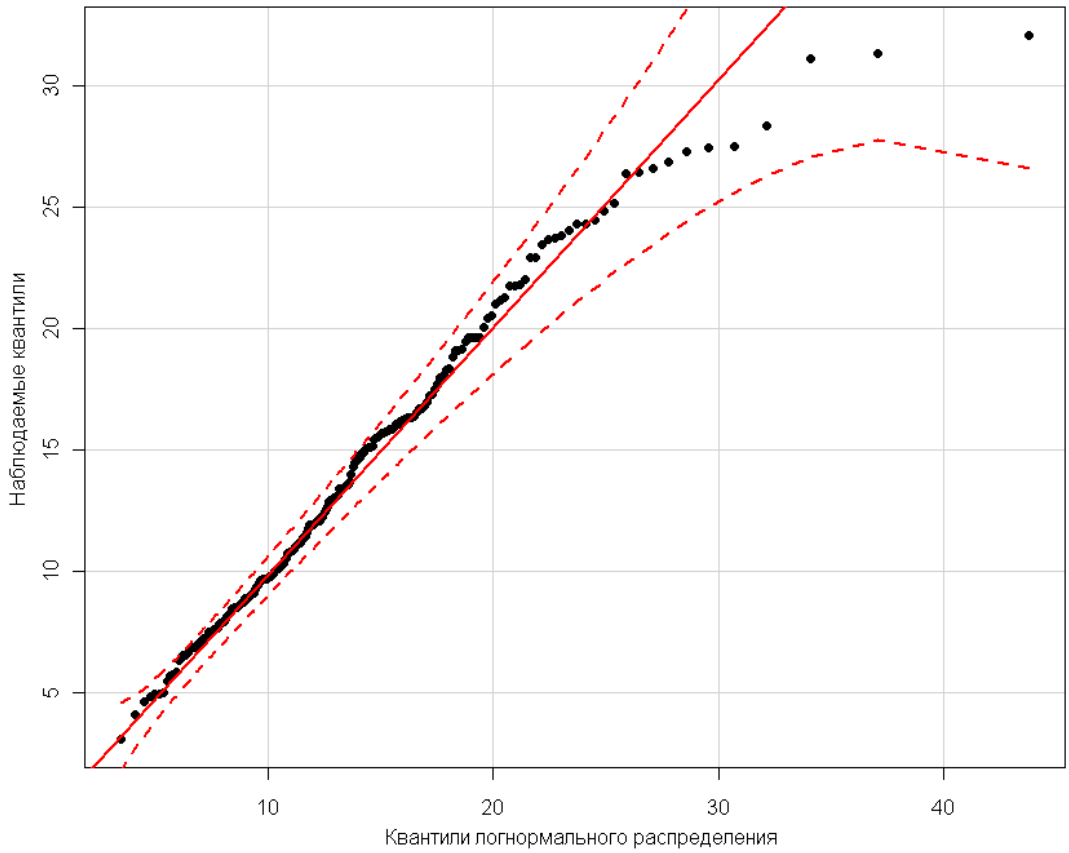


Рис. 1.11. 95%-е доверительные огибающие при аппроксимации выборочной массы тела ящериц логнормальным распределением

Рассмотрим еще один пример. Вильфредо Парето, известный экономист конца XIX века, политик и прото-фашист, разработал закон неравенства богатства, который лег в основу семейства гиперболических распределений, играющих важнейшую роль в экологии и других науках о сообществах (см. законы Ципфа, Уиллиса, Фишера, Лотки, Лоренца – Левич, 1978; Шитиков и др., 2012). Распределение Парето (Clauset et al., 2007) или степенная функция описывает универсальную модель данных с “мощным хвостом”, то есть когда плотность вероятности $p(x)$ очень медленно сходится к нулю при $x \to \infty$:

$$p(x) = \frac{\theta - 1}{x_0}\left(\frac{x}{x_0}\right)^{\theta}; \quad \hat{\theta} = 1 + \frac{n}{\sum_{i=1}^{n} \log x_i / x_0}$$

где θ – показатель степени, x_0 – минимальное начальное значение. Это распределение имеет сильную правостороннюю асимметрию со средним, значительно превышающим медиану.

Выполним аппроксимацию распределением Парето ряд средних численностей $\{N_1, N_2, \ldots, N_{188}\}$ 188 видов донных организмов по результатам взятия 55 гидробиологических проб из реки Байтуган [пример П2]. Воспользуемся для этого набором функций из файла pareto.r, разработанных проф. К. Шализи (Shalizi, http://www.stat.cmu.edu/~cshalizi/) и осуществляющих вычисление параметров распределения θ и x_0 на основе метода максимального правдоподобия.

График на рис. 1.12 показывает распределение $F(N)$ популяционных плотностей видов макрозообентоса в логарифмических шкалах с оцененным показателем степени $\hat{\theta} = 1.61$, причем наилучшая аппроксимация учитывает только 82 вида с минимальной численностью особей $x_0 = 160$.

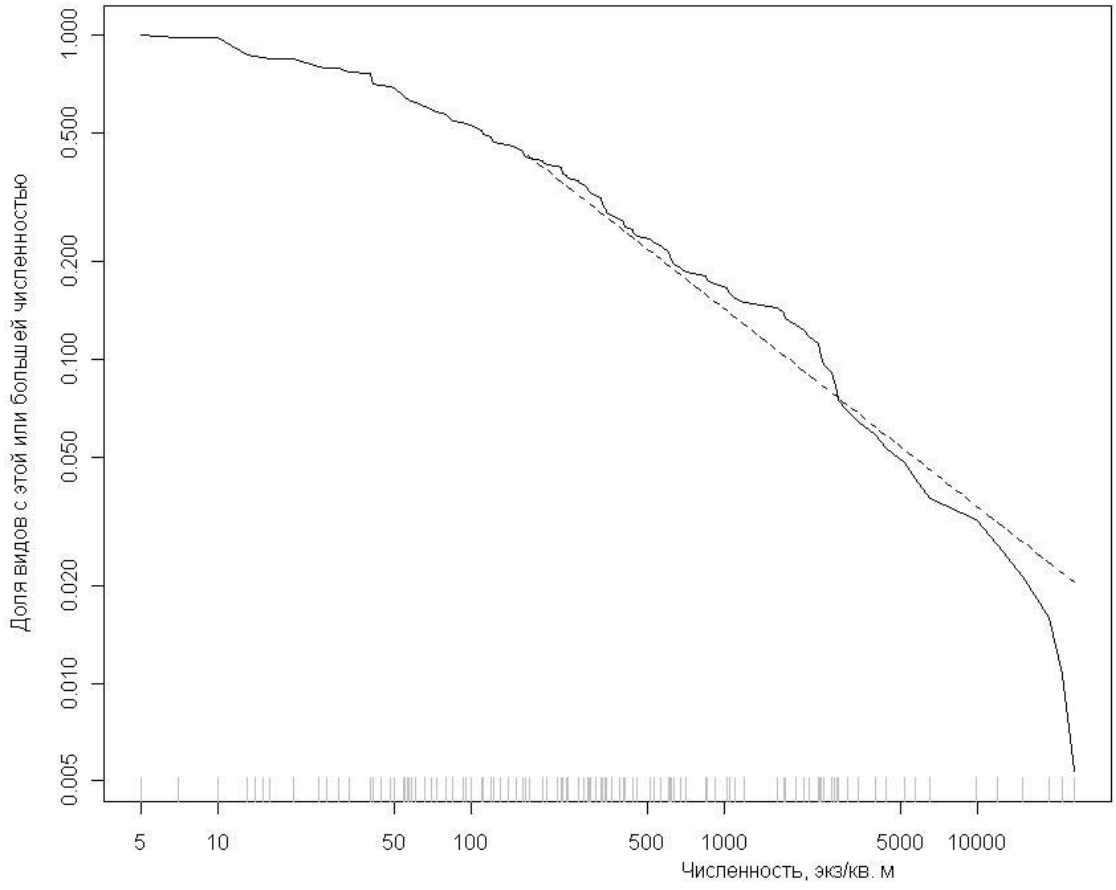


Рис. 1.12. Кумулятивные функции распределения численности видов донных организмов. Сплошная линия показывает эмпирическую долю количества видов из 188, численность которых

превышает значение по оси абсцисс. Пунктир – теоретическое распределение Парето при $x_0 = 160$, вычисленное методом максимального правдоподобия. Шкалы логарифмированы.

Зададимся вопросом, насколько можно доверять найденным оценкам параметров. Для этого многократно (B = 1000) будем извлекать из теоретического распределения Парето ($\theta = 1.61$, x_0 = 160) псевдовыборки размером n = 188 и для каждой из них вычислять значения $\hat{\theta}^*$. На основе этого можно получить по формуле (1.5) бутстреп-оценки стандартной ошибки, смещения и основных доверительных интервалов (1.9) для показателя степени $\hat{\theta}$:

$$se_\theta = 0.0468; \quad |\hat{\theta} - \bar{\hat{\theta}}^*| = 0.00345; \quad l_н = 1.51; \quad l_в = 1.69.$$

Теоретически те же доверительные интервалы могли бы быть рассчитаны параметрическим методом с использованием обратного γ-распределения, но это сложнее и требует серьезной методической проработки, чем просто выполнить бутстреп.

Другая краеугольная проблема нашего примера – это оценить, насколько хорошо эмпирические данные, представленные на рис. 1.12, согласуются с предполагаемым теоретическим распределением. Пусть нулевая гипотеза H_0 гласит, что выборка численностей бентосных организмов подчиняется закону Парето, и если p-значение, соответствующее статистике Колмогорова-Смирнова D, будет меньше 0.05, то она отклоняется в пользу альтернативной гипотезы. К сожалению, далеко не всегда асимптотические процедуры оценки p-значений оказываются корректными (например, не допускается наличие повторяющихся значений). Поэтому на минуту забудем о существовании таблиц критических значений D_α и обратимся за помощью к процедурам ресамплинга.

Бутстреп дает возможность не только рассчитывать интервальные значения параметров, но и проверять статистические гипотезы, о чем подробно пойдет речь в следующей главе. В общем случае при проверке гипотез интерес представляют два набора выборочных распределений произвольной тестовой статистики T: а) распределение при справедливости нулевой гипотезы, которое дает возможность градуировать T по уровням значимости, и б) распределение под альтернативой, позволяющее оценить достигнутую мощность. Если речь идет об эмпирических выборках случайных величин, то оба этих распределения T мы можем легко сформировать бутстрепом, что и будет показано в дальнейшем. Однако если мы попытаемся оценить таким образом согласие параметров модели, рассчитанной по конкретной выборке, с параметрами теоретического распределения, то мы сталкиваемся с тем, что распределение тестового критерия будет урезанным. Т.е. оно не может считаться «полученным при справедливости нулевой гипотезы», поскольку оцениваемые параметры были уже изначально оптимизированы под исходный набор эмпирических данных.

Разрешить эту проблему нам поможет механизм "двойного бутстрепа", который для рассматриваемого примера реализуется следующим образом:

- вычисляется статистика Колмогорова-Смирнова D_{obs} для эмпирических данных;
- рассчитывается множество (B = 1000) значений статистики D^*_{sim} для моделей, полученных на независимых псевдо-выборках, сформированных во "внутренней петле" бутстрепа;
- находится число случаев b, когда имело место $D^*_{sim} > D_{obs}$, и вероятность ошибки 1-го рода будет равна $p = (b + 1)/(B + 1)$.

На "внутренней петле" из распределения Парето с фиксированными параметрами $\hat{\theta}$ и x_0 сначала извлекается n случайных значений, по которым рассчитываются новые параметры $\hat{\theta}^*$ и x_0^* нуль-модели, зависящей от исходных данных уже только через оценку $\hat{\theta}$.

Для рассматриваемого примера мы получили p = 0.016, т.е. несоответствие между эмпирическими численностями особей и предсказанными по закону Парето настолько велико, что случайно это может иметь место лишь в двух случаях из 100. В то же время отметим, что, даже посчитав модель Парето недостаточно правильной, все равно

полученная нами оценка $\hat{\theta}$ будет сходиться к значению, которое является, в определенном смысле, наилучшим приближением параметра истинного распределения в классе большинства степенных функций (Clauset et al., 2007).

К разделу 1.5:

```
#    а)  Бутстрепирование медианы и метод обратной трансформации
#  Определение вектора x с выборочными данными численности личинок мотыля
x <-c(5, 5, 8, 10, 10, 10, 19, 20, 20, 20, 30, 40, 42, 50, 50, 50, 50, 50, 50, 65, 65, 80,
80, 80, 100, 100, 133, 200, 250, 300, 430, 440, 480, 800, 880, 2400, 3020, 3360, 5200, 6200,
7000, 9000, 19000)
#  Расчет бутстреп-распределения медианы для исходной выборки
library(boot) ; f.Med <- function(y,id) {median(y[id])} # Функция, определяющая статистику
# ------------------------ Генерация псевдочисленностей методом обратной трансформации
# Определим предварительно функцию, возвращающую таблицу,
# содержащую Nmed строк, где Nmed - заданное число псевдо-медиан,
# сгенерированных из функции эмпирического распределения X
Genmed <- function (x, Nmed) {
       # Формируем список уникальных значений X и число полученных градаций
       n <-length(x) ; xg <- unique(sort(x)) ; ng <- length(xg)
       # и вектор накопленных относительных частот
       p <- cumsum(as.data.frame(table(sort(x)))$Freq/n)
       Meds <- numeric(Nmed) ; for (j in 1:Nmed) {
       # Получаем случайное число (0 < tp <1) и находим диапазон частот, где оно находится
       tp <-  runif(1); i <- length(which(p <= tp, arr.ind = FALSE))
       #  Вычисляем  псевдочисленность простой линейной интерполяцией
       if (i == 0) { xtp = xg[1]*tp/p[1] } ; if (i == ng) { xtp = xg[ng] }
       if  (i > 0  &  i < ng ) { xtp = xg[i]+(xg[i+1]- xg[i])*(tp - p[i])/(p[i+1]-p[i]) }
       Meds[j] <- xtp } ; return (Meds) }
# ------------------------ Выполнение расчетов
Med200 <- Genmed (x, 200) ; quantile(Med200) ; plot(density(Med200),col="black", lwd=2)
bootres <- boot(Med200, f.Med, 5000) # накопление результатов в Bootobject
# вывод комплекта из различных версий доверительных интервалов
boot.ci(bootres, conf = 0.95, type = c("norm", "basic", "perc", "bca"))
#    б) Параметрический бутстреп
#  Функция оценки параметра λ для выборочного распределения Пуассона
lambda_MLE <- function (data) as.numeric(fitdistr(data, "Poisson")$estimate
#  Функция формирования бутстреп-распределения  оцениваемого параметра
lambda_boot <- function (data, lambda, B=2000, param = TRUE) {
  result  <- rep(0,B) ; n <- length(data)
  for(i in 1:B)   {
      if (param)  y.boot <- rpois(n, lambda)   #   Параметрический бутстреп
      else   y.boot <- sample(data, n, replace=T) #   Непараметрический бутстреп
      result[i] <- lambda_MLE(y.boot) }
   return (result) }
x <-c(rep(0,450),x)  #   Добавление нулевого хвоста из 450 значений
lambda_Bootstrap <- lambda_boot (x, lambda_MLE(x))
lambda_Bootstrap <- lambda_boot (x, lambda_MLE(x), param = FALSE)
#    Отрисовка графика на рис. 1.8
hist(lambda_Bootstrap, xlab = "Значение параметра", breaks=50, prob=TRUE)
curve(dnorm(x, mean=mean(lambda_Bootstrap),
                    d=sd(lambda_Bootstrap)),col='blue',add=TRUE,lwd=2)
# Вычисление верхнего и нижнего доверительных интервалов
delta <-quantile((lambda_Bootstrap  - lambda_MLE(x)), probs=c(0.025,0.975), names=FALSE)
L <- lambda_MLE(x)- delta [2] ; abline(v= L, col = "green",lwd=2)
U <- lambda_MLE(x) - delta [1] ; abline(v= U, col = "red",lwd=2)
#    в) подбор распределений
# Загрузка выборочных данных по массе тела ящерицы Zootoca из текстового файла
Z <- read.delim("Zootoca.txt") ; x = Z$bm ; summary(x); n <- length(x)
# Построение доверительных огибающих (confidence envelope)
library(boot) ; z <- (x - mean(x))/sqrt(var(x))  #  Стандартизация выборки
x.qq <- qqnorm(z, plot.it = FALSE) ; x.qq <- lapply(x.qq, sort)
plot(x.qq, ylim = c(-2, 5), ylab = "Выборочные Z-значения",
                            xlab = " Z-значения для нормального распределения")
```

```
# Генерация 999 бутстреп-выборок (т.е. случайных выборок из нормального
#  распределения с установленными оценками параметров для выборки z)
x.gen <- function(dat, mle) rnorm(length(dat))
x.qqboot <- boot(z, sort, R = 999, sim = "parametric",ran.gen = x.gen)
# Определяем и рисуем огибающие
x.env <- envelope(x.qqboot, level = 0.9)
lines(x.qq$x, x.env$point[1, ], lty = 4, lwd=2)
lines(x.qq$x, x.env$point[2, ], lty = 4, lwd=2)
lines(x.qq$x, x.env$overall[1, ], lty = 1, lwd=2)
lines(x.qq$x, x.env$overall[2, ], lty = 1, lwd=2)
# Функции отрисовки графиков ЭФР  и ЭФПР
graph_ECFD <- function (x, pc)
       #   Кумулятивные функции распределения
       { plot(ecdf(x)) ; lines (x,pc, type="l",col="red", lwd=2) }
graph_EFDD <- function (x, pd)
#   Функции плотности распределения (ядерная и теоретическая)
{ plot(density(x), lwd = 2, col="blue", ylim=c(0,0.082)) ; lines(x, pd,  col="red",lwd = 2)}
library(MASS) ; library(car) ##  оценка параметров распределений
##  оценка параметров нормального распределения
(dof = fitdistr(x,"normal")) ; ep1=dof$estimate[1]; ep2=dof$estimate[2]
ks.test(x,pnorm, mean=ep1,sd=ep2)
##  оценка параметров распределения Вейбулла
(dof = fitdistr(x,"weibull",start=list(scale=1,shape=2)))
ep1=dof$estimate[1]; ep2=dof$estimate[2]
ks.test(x,pweibull, scale=ep1,shape=ep2)
##  оценка параметров логнормального распределения
(dof = fitdistr(x,"log-normal")) ; ep1=dof$estimate[1]; ep2=dof$estimate[2]
ks.test(x,plnorm, meanlog=ep1,sdlog=ep2)
graph_ECFD (x, plnorm(x,meanlog=ep1,sdlog=ep2))
graph_EFDD (x, dlnorm(x,meanlog=ep1,sdlog=ep2))
qqPlot(x, dist= "lnorm",meanlog=ep1,sdlog=ep2,xlab="Квантили логнормального распределения",
                                              ylab="Наблюдаемые квантили", pch=19)
###   г)  Моделирование распределения видовой численности сообщества функцией Парето
#  Загрузка данных (численности и биомассы видов) из текстового файла с разделителями
TB <- read.table("ABC_13.txt",header=T,row.names=1,sep="\t"); n=length(TT)
TT <- t(TB[,1]);  TT <- rev(sort(TT[TT>0]))  # Транспонирование вектора численносей по видов
source("pareto.R") # Загрузка комплекта функций, связанных с распределением Парето
TT.pareto <- pareto.fit(TT,threshold="find") #  Оценка параметров модели
X0 <- TT.pareto$xmin ; teta <-TT.pareto$exponent  ; D <- TT.pareto$samples.over.threshold
#  Вывод графика эмпирических данных и теоретического распределения
plot.survival.loglog(TT,xlab="Численность, экз/кв. м", ylab="Доля числа видов ")
rug(TT,side=1,col="grey")
curve((D/n)*ppareto(x,threshold=X0,exponent=teta,lower.tail=FALSE),
             add=TRUE,lty=2,from=X0,to=max(TT))
# Параметрический бутстреп для оценки характеристик параметра teta распределения Парето
# Функция генерирует B наборов данных размером n из распределения Парето с параметрами
# exponent и x0 и возвращает вектор из B значений пересчитанных показателей степени teta
rboot.pareto <- function(B,exponent,x0,n) {
replicate(B,pareto.fit(rpareto(n,x0,exponent),x0)$exponent)}
tboot <- rboot.pareto(1000,teta,X0,n) ; alpha = 0.05  # Выполнение расчетов
sd(tboot) ;  (mean(tboot) - teta)  # стандартная ошибка и смещение
# 95% доверительные интервалы  по методу Халла
(ci.lower <- 2*teta - quantile(tboot,1-alpha/2)) ; (ci.upper <- 2*teta -
quantile(tboot,alpha/2))
# Параметрический бутстреп для оценки p-значения теста Колмогорова-Смирнова
# Вычисление статистики Колмогорова-Смирнова для части данных
ks.stat.pareto <- function(data, exponent, x0) {
       ks.test(data[data>=x0], ppareto, exponent=exponent, threshold=x0)$statistic}
#  Вычисление p-значения для статистики Колмогорова-Смирнова
ks.pvalue.pareto <- function(B, data, exponent, x0) {
       testhat <- ks.stat.pareto(data, exponent, x0) ;   testboot <- vector(length=B)
  for (i in 1:B) {    xboot <- rpareto(length(data),exponent=exponent,threshold=x0)
             exp.boot <- pareto.fit(xboot,threshold=x0)$exponent
             testboot[i] <- ks.stat.pareto(xboot,exp.boot,x0) }
```

```
  p <- (sum(testboot >= testhat)+1)/(B+1)  return(p) }
# ---------------------- Выполнение расчетов
ks.stat.pareto(TT,teta,X0) ; ks.pvalue.pareto(1000,TT,teta,X0)
```

1.6. Бутстрепирование индексов, характеризующих многовидовые композиции

Большинство экологических, социальных или экономических систем представляют собой композиции объектов, принадлежащих многим видам. Таковы, например, сообщество организмов, населяющих дно водоемов, номенклатура электромоторов, выпускаемых предприятием, совокупность товаров, которыми торгует супермаркет, профессиональный состав жителей микрорайона и т.д. Отдельные комплексы видов могут сравниваться между собой по уровню разнообразия, характеру распределения общей численности по отдельным компонентам, мере сходства сообществ в пространстве видов и др. Для этого разработано большое количество разнообразных показателей, которые часто называют *индексами*.

Каждый вид количественно характеризуется относительной частотой его встречаемости $p_i = N_i / \sum_{i=1}^{S} N_i$, где N_i – число экземпляров *i*-го вида из их общего числа *S*. Иногда в эту формулу вместо численностей подставляют суммарную массу, стоимость или площадь, соответствующую каждому виду. Для оценки видового разнообразия используются два основных показателя:

- индекс *Джини-Симпсона* $C = \sum_{i=1}^{S} p_i^2$, который является смещенной оценкой дисперсии p_i;
- информационный индекс *Шеннона* $H = \sum_{i=1}^{S} p_i \log_2 p_i$, соответствующий среднему минимальному числу испытаний, в результате которых из сообщества будет извлечен экземпляр самого многочисленного вида (или в нитах $H = \sum_{i=1}^{S} p_i \ln p_i$).

Модель распределения общей численности *N* сообщества по видам можно представить, например, сходящимся логарифмическим рядом Маклорена: αx, $\alpha x^2/2$, …, $\alpha x^n/n$, где $\alpha x^i/i$ – количество видов в группе с *i*-й численностью экземпляров, $x = N/(N - \alpha)$. Параметр α логсерии, известный как α *Фишера*, показывает, как круто падает кривая распределения частот в ранжированном ряду видов, начиная с самого многочисленного, т.е. также является индексом выравненности сообществ.

Пусть, например, в результате взятия 147 гидробиологических проб в средней равнинной реке Сок с его притоке реке Байтуган [пример П2] было обнаружено 375 видов и таксономических групп макрозообентоса. Поставим задачу установить, имеются ли различия в пределах трех укрупненных участков (Байтуган, верховья реки Сок и ее нижний участок), или вся речная система является однородной по уровню биоразнообразия донных сообществ. Для каждого из трех участков были рассчитаны средние численности особей бентоса по всем взятым пробам, которые затем округлялись до целочисленного значения (табл. 1.2). Несмотря на то, что сам состав видов между участками значительно различался, значения показателей общего видового разнообразия, оцениваемого по вышеперечисленным индексам, оказались весьма близкими. Однако поставим своей целью подтвердить это предположение статистическими методами.

Ранговое распределение видов на каждом участке имеет традиционный гиперболический характер с очень длинным хвостом из редких и малочисленных таксономических групп. Наблюдаемые численности настолько хорошо аппроксимируются моделями Ципфа и Мандельброта (рис. 1.13), что значения индекса Шеннона, рассчитанные по эмпирическим и модельным численностям, весьма незначительно отличаются между собой. Эти модели можно использовать для оценки

доверительных интервалов индексов разнообразия с использованием функций максимального правдоподобия (Chao, Shen, 2003).

Таблица 1.2. Средние численности видов макрозообентоса и показатели видового разнообразия для трех участков речной системы (фрагмент)

пп	Виды зообентоса	Байтуган	Сок_верх	Сок_нижн	Итого
1	*Eukiefferiella* gr.gracei	999	124	0	1123
2	*Tanytarsus* sp.	132	542	140	814
3	*Paracladius conversus*	274	495	18	787
4	*Cladotanytarsus mancus*	0	193	353	546
5	*Chironomus nudiventris*	0	0	542	542
6	*Lipiniella araenicola*	0	0	524	524
7	*Prodiamesa olivacea*	118	357	0	475
8	*Baetis rhodani*	435	13	3	451
9	*Cricotopus bicinctus*	12	360	49	421
308	*Rhyacophila* sp.	0	1	0	1
309	*A. piscinalis*	0	0	1	1
310	*Anodonta* sp.	0	0	1	1
	Итого	4465	4954	4720	14139
	Индекс Шеннона *H*	3.27	3.20	3.29	
	Индекс Симпсона *C*	0.9157	0.9181	0.9227	
	α Фишера	30.43	30.15	27.58	

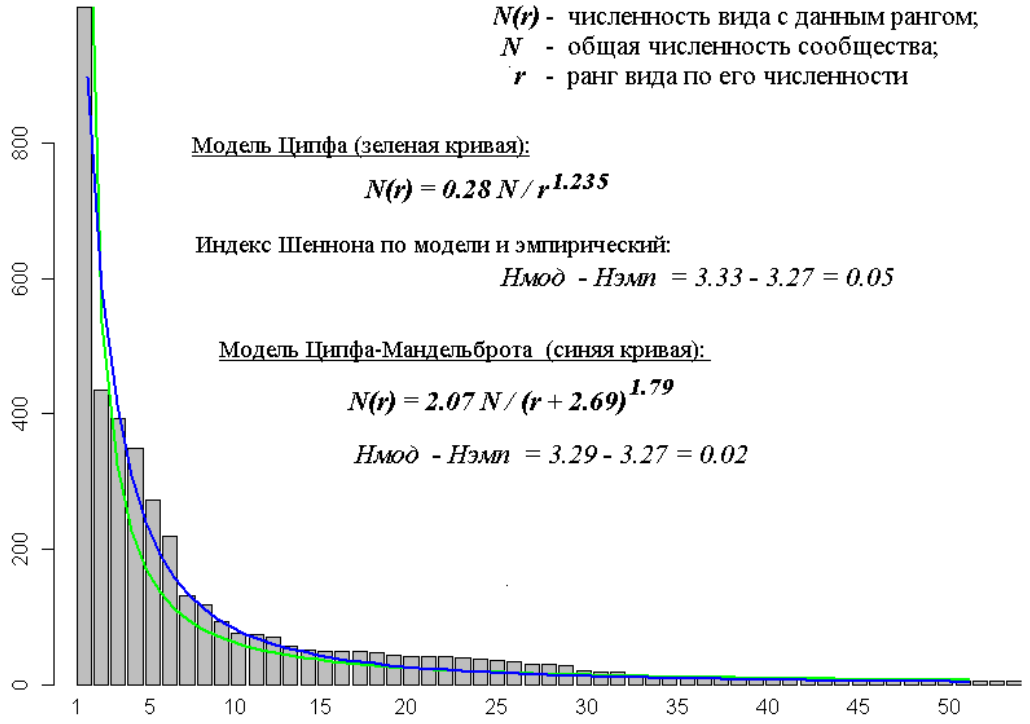


Рис. 1.13. Модели Ципфа-Мандельброта распределения по видам общей численности донного сообщества в р. Байтуган (представлена левая треть шкалы численностей)

Стандартную ошибку и доверительные интервалы произвольного индекса разнообразия можно также оценить с использованием различных методик бутстрепа. При использовании классического непараметрического бутстрепа на основе алгоритма "случайного выбора с возвращением" из исходного вектора численностей видов $\{N_1, N_2, \ldots, N_S\}$ формируется большое количество ($B = 5000$) псевдовыборок той же размерности. Некоторые виды в бутстреп-выборке могут отсутствовать, в то время как другие значения

исходных численностей могут появляться в ней дважды или более раз. Для каждой из смоделированных композиций видов рассчитывается индекс разнообразия, строится функция его распределения (рис. 1.14), после чего находятся бутстреп-оценки смещения среднего, стандартной ошибки индекса (формула 1.7) и доверительные интервалы по методу процентилей (1.8) или с использованием иных формул.

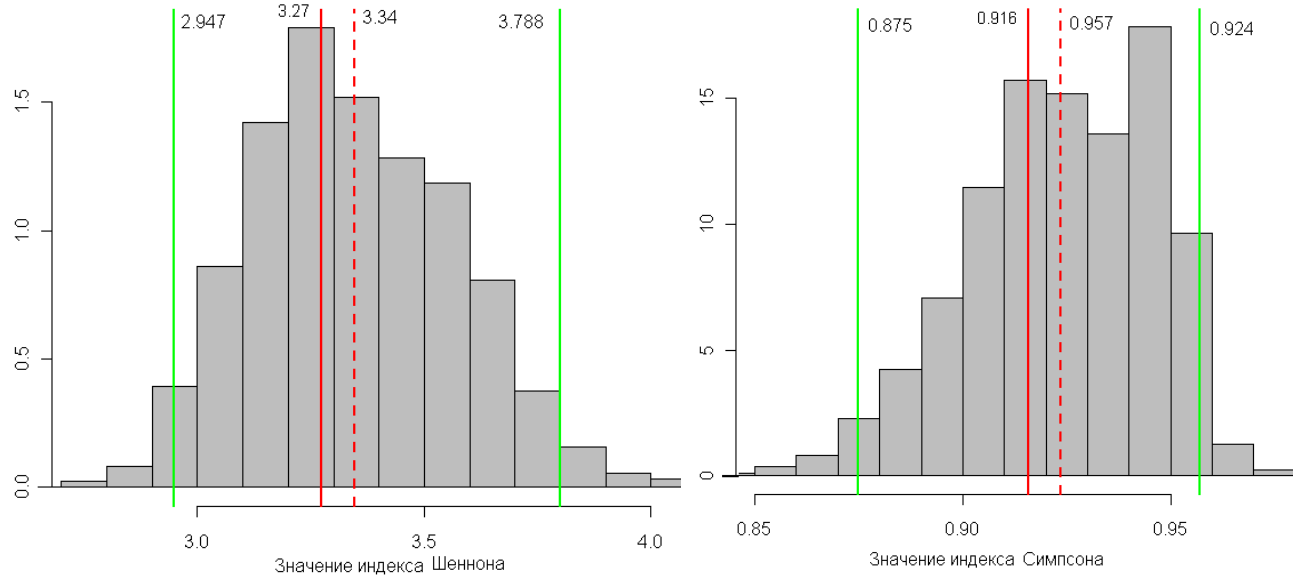


Рис. 1.14. Гистограммы частотного распределения 5000 бутстреп-значений индексов Шеннона (а) и Симпсона (б); зеленые линии – нижняя и верхняя границы доверительного интервала, найденные методом процентилей, красная сплошная линия – эмпирические значения, пунктирная линия – среднее значение бутстрепа.

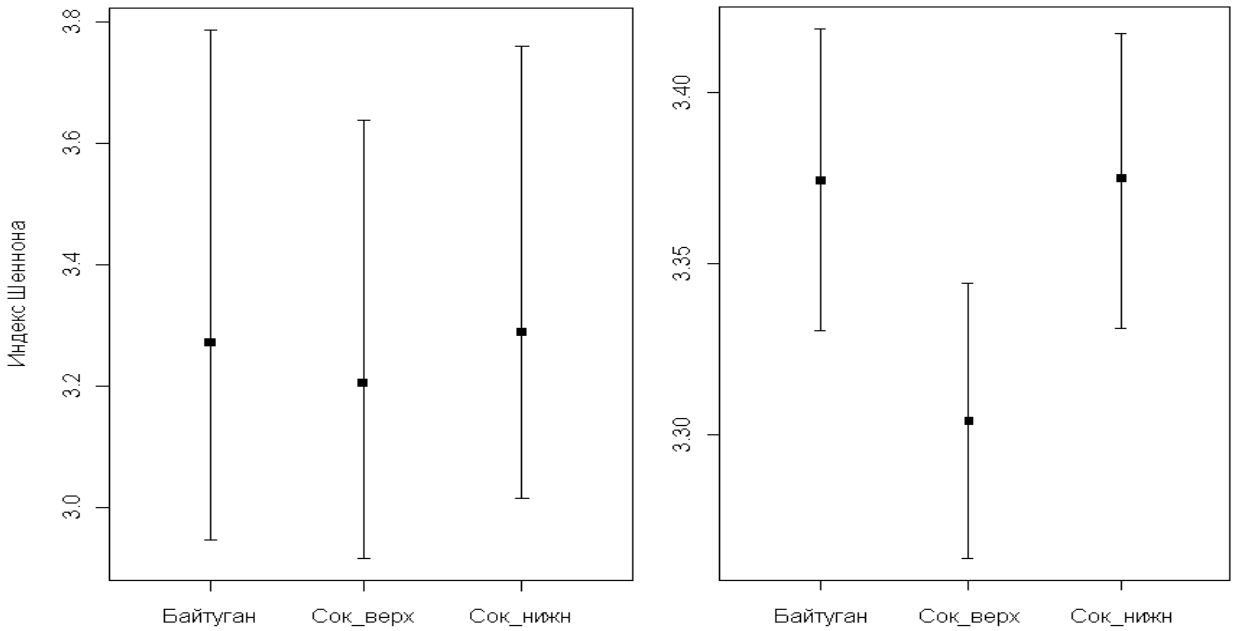
Представленный алгоритм реализован во многих программах (например, PAST), но, на наш взгляд, при большой гетерогенности численностей ширина доверительного интервала на рис. 1.14 является несколько завышенной. Основанием для сомнений является то обстоятельство, что в ходе случайного выбора в составе бутстреп-выборки могут исчезать некоторые доминирующие виды, и это будет коренным образом изменять ситуацию с оценками биоразнообразия (тем самым, расширять доверительные пределы).

Сформулируем два таких, на первый взгляд, разумных ограничения, регулирующих свободу комбинаторики в процессе перебора вариантов бутстрепа: а) элементы псевдовыборок не обязательно должны в точности совпадать с величинами из исходной совокупности, но генерироваться на основе ее эмпирической функции распределения; б) любая из полученных выборок должна иметь одну и ту же суммарную численность особей. Рассмотрим другой алгоритм непараметрического бутстрепа, предложенный на блоге Кея Сичини (Kay Cichini) и генерирующий значения численностей из распределения, асимптотически приближенного к эмпирическому.

Пусть донное сообщество реки Байтуган на каждом кв. метре состоит из 4465 особей, принадлежащих 152 видам. Для получения бутстреп-выборки будем последовательно извлекать каждую особь и случайным образом относить к одному из этих видов. Если мы потом подсчитаем численности каждого вида, то при равновероятном присваивании все они будут близки к 30. Теперь введем веса, равные вероятностям встречаемости каждого вида в исходной выборке. Тогда случайное назначение с учетом этих весов даст нам псевдо-выборку с тем же общим числом особей, но их распределение по видам будет лишь немного отличаться от эмпирического.

Бутстреп-ошибка и ширина доверительных интервалов индекса, оцененные по второму алгоритму, оказались примерно в десять раз меньше, что может рассматриваться как некая вполне обоснованная объективность (рис. 1.15). К счастью, в нашем случае

различие в алгоритмах бутстрепирования не отразилось на результатах сопоставления показателей разнообразия для различных изучаемых объектов. Так доверительные интервалы индекса Шеннона для различных участков р. Сок и его притока р. Байтуган, во всех комбинациях попарно пересекаются, поэтому можно говорить, что в этой системе водотоков имеются однородные по видовому разнообразию донные сообщества.



а) б)

Рис. 1.15. Диаграмма оценки видового разнообразия макробентоса различных участков р. Сок на основе доверительных интервалов, построенных двумя алгоритмами бутстрепа: случайным выбором с возвращением (а) и на основе эмпирической функции распределения численностей (б)

Здесь мы коснулись важнейшей методологической проблемы ресамплинга: выбора разумных ограничений, контролирующих "вольность" генерации псевдовыборок. Если эти ограничения будут излишне жесткими, то доверительные интервалы сужаются и возникает потенциальный риск ошибки 1-го рода. Негативные явления противоположного свойства имеют место при полной бесконтрольности перебора и тогда возрастает риск ошибки 2-го рода, т.е. принять нулевую гипотезу, когда она неверна. Рецепт, как счастливо избежать этих обеих опасностей, в настоящее время отсутствует и будет, вероятно, предметом долгих обсуждений математиков-специалистов и прикладных исследователей в ближайшие десятилетия.

Один из способов получить представление о видовой структуре сообщества состоит в анализе кумулятивной кривой – графика, на котором по оси абсцисс откладываются порядковые номера видов, ранжированные по их численности, а по оси ординат – накопленные значения p_i. Как и на графиках Лоренца, чем больше отклонение эмпирической кривой от главной диагонали, которая совпадает с линией равного распределения долей (или максимальной "эквитабельности"), тем в большей степени виды в сообществе представлены неравномерно. Р.Уорвик и К. Кларк (Warwick, Clarke, 1994), изучавшие морской бентос, на одном графике размещали сразу две кумуляты: для численности организмов и их биомассы, и оценивали площадь между этими кривыми. В рамках разработанного ими ABC-метода (Abundance/Biomass Comparisons) рассчитывается W-статистика Кларка:

$$W = \sum_{i=1}^{S} (Bc_i - Nc_i) / 50(S-1),$$

где Bc_i и Nc_i – накопленные относительные значения биомассы и численности (%) для i-го по рангу вида; S – число видов. Если $W > 0$, то кумулятивная кривая биомассы

располагается выше кривой численности, что является признаком устойчиво развивающегося сообщества.

Ниже представлен фрагмент списка видов макрозообентоса для расчета величины W по результатам гидробиологической съемки на р. Байтуган:

Виды	N_i, экз.	Nc_i, %	B_i, г.	Bc_i, %	Bc_i - Nc_i	
1. *Eukiefferiella* gr.*gracei*	54934	21.98	4.15	84.85	62.87	
2. *Baetis rhodani*	23914	31.62	19.72	66.20	34.57	
3. *Limnodrilus profundicola*	21670	40.29	81.07	19.07	-21.22	
4. *Ephemerella ignita*	19202	47.98	38.52	54.62	6.6	
…						
186. *Rheotanytarsus curtistylus*	10	99.996	0.01	99.97	-0.02	
187. *Pisidium* sp.	5	99.998	0.01	99.92	-0.069	
188. *Cricotopus albiforceps*	5	100	0.01	99.93	-0.062	
				Итого	281.51	$W = 0.0301$

Бутстреп-процедура для оценки доверительных интервалов W-статистики мало отличается от представленной выше схемы анализа индексов разнообразия. Здесь только одна тонкость – бутстепированию достаточно подвергнуть не всю таблицу наблюдений, а только вектор разностей (Bc_i - Nc_i). В результате легко получит стандартную ошибку статистики Кларка $s_W = 0.0137$ и ее доверительные интервалы $CI_{95\%} = 0.0029 \div 0.057$.

Основная задача гидробиологии – выделить подмножество видов, которые являются индикаторами загрязнения воды различного типа. Было отмечено, что органическое загрязнение сопровождается ростом доли малощетинковых червей подкласса Oligochaeta (за исключением рода Nais). Отношение их численности к численности всего бентоса (%) было названо олигохетным индексом Гуднайта-Уитли J.

В качестве исходных данных используем таблицу численностей 120 видов по результатам наблюдений на устьевом участке р. Сок, из которых только 5 относятся к подклассу олигохет. Но и при этом основная масса малощетинковых червей принадлежала только одному из видов тубифицид: 35000 экз. *Tubifex tubifex* против 1500 экз., принадлежащих остальным 4-м видам. В ходе бутстрепа все псевдовыборки разделились на две части: включающие вид *Tubifex tubifex* с большой численностью и композиции видов с его отсутствием, для которых индекс J был близок к нулю – см. рис. 1.16.

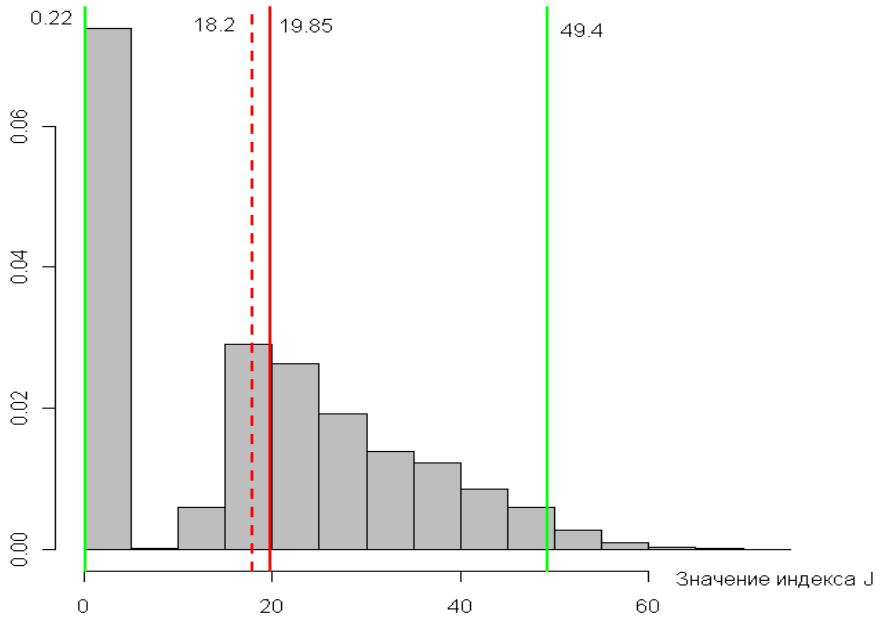
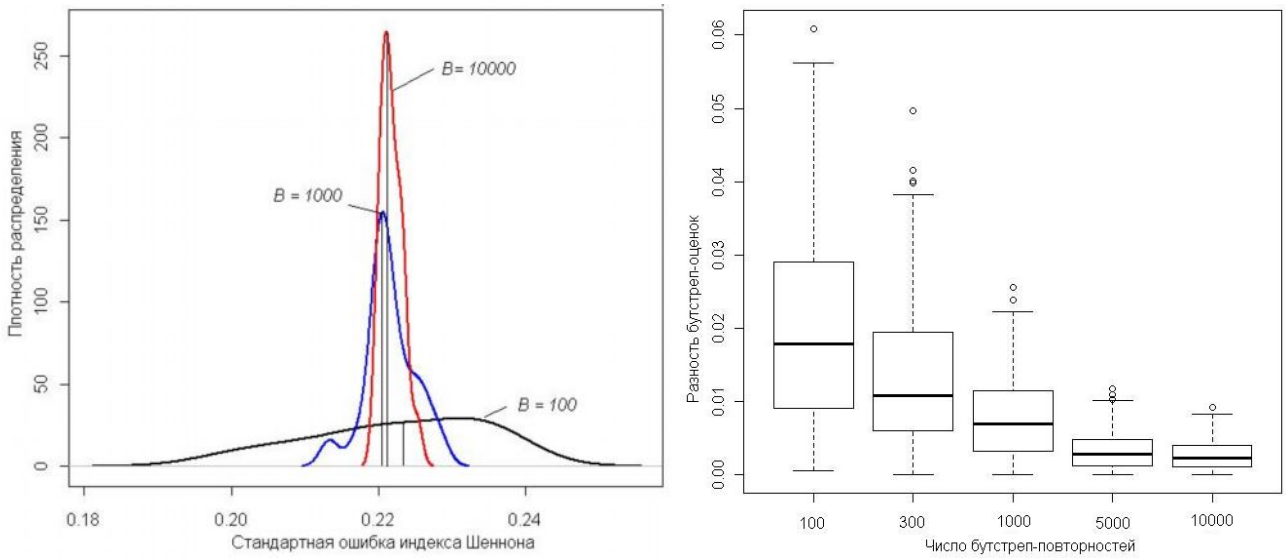


Рис. 1.16. Гистограмма частотного распределения 5000 бутстрепированных значений индекса Гуднайта-Уитли (обозначения те же, что и на рис 1.14)

Результаты оценки доверительных интервалов олигохетного индекса оказались не слишком впечатляющими, если не сказать безобразными. Это свидетельствует о том, что в любом случае использование бутстрепа должно основываться на предварительном изучении конкретного распределения исходных данных.

При выполнении бутстреп-процедур возникает другой естественный вопрос: насколько большим должно быть задаваемое число повторностей *B*, чтобы обеспечить разумную точность оценки параметров? В общем случае при увеличении *B* процесс оценивания монотонно и достаточно быстро сходится, хотя на небольших выборках и при достаточной производительности компьютера число итераций можно задавать десятками миллионов. Но есть ли в этом объективная необходимость?

Зададимся последовательностью значений *B* = {100, 300, 1000, 5000, 10000} и для каждого числа бутстреп-итераций выполним по 20 повторностей расчета индекса Шеннона *H* в р. Байтуган. Результаты вычислений представим графиками на рис. 1.17.



а) б)

Рис. 1.17. Функции плотности распределения стандартной ошибки индекса Шеннона (а) и абсолютное смещение бутстреп-оценки относительно своего наилучшего результата (б)

Обратим внимание на то, что уже после 100 итераций мы получаем вполне статистически состоятельные бутстреп-оценки: значения индекса *H* и его ошибки только в третьей значащей цифре отличаются от полученных при *B* = 10000. Однако при таком малом числе перевыборок возникает опасность случайно получить в некотором смысле аномальные значения, основанные на нехарактерных (экстремальных) комбинациях элементов исходной выборки. При увеличении итераций такая опасность исчезает, а диапазон возможных значений бутстреп-оценок достаточно быстро сужается. Например, в нашем случае минимальные и максимальные значения *H* из 20 повторностей при *B* = 100 итераций находились в пределах 3.328 ÷3.378, а при *B* = 10000 этот интервал составил уже 3.344 ÷ 3.353. Разумеется, здесь многое может зависеть от структуры эмпирических данных, поэтому окончательные рекомендации обычно даются с известной осторожностью.

К разделу 1.6:

```
#   Бутстреп индексов разнообразия и биотических индексов
library(vegan) # Загрузка пакета vegan с функциями расчета биоразнообразия
source("print_rezult.r")  # Загрузка функций вывода результатов
# --------------------
# Определение функции, осуществляющей расчет индексов разнообразия
#  Параметры:  indName - идентификатор индекса; data - вектор численностей
indiv <- function (indName, data) {
```

```
   if (indName=="shannon") ind <- diversity(data)              #  Индекс Шеннона
   if (indName=="simpson") ind <- diversity(data,"simpson")    # Индекс Симпсона
   if (indName=="alpha") ind <- fisher.alpha(data)                  # Альфа логсерий Фишера
   return (ind) }
# Определение функций, выполняющих формирование бутстреп-выборок индексов разнообразия
#  Параметры: method - алгоритм перевыборки, permutations - число итераций бутстрепа
ici <- function(indName, data , method = 1, permutations=5000,Hist=FALSE)
   { boots <- numeric(permutations) ; data <- data[data>0] # Исключение численностей = 0
    results <- numeric(4) ;   for (i in 1:length(boots)) {
## -----  Непараметрический бутстреп случайным выбором с возвращением
    if (method == 1) boots[i] <- indiv(indName, sample(data,replace=T))
## ---  Бустреп-процедура, генерирующая значения из  распределения,
##       асимптотически приближенного к эмпирическому
    else { spn<-length(data) ; obs<-sum(data) # Суммарная численность
   probs<-data/obs   # Вектор весов для создания распределения
   temp.counts<-rep(1,spn)+ tabulate(sample(1:spn,obs-spn,prob=probs,replace=T),nbins=spn)
   boots[i] <- indiv(indName, temp.counts)}
  }
div = indiv(indName, data) ; bias=mean(boots)-div ; results[1]<-div ; results[2]<-bias
CI <- quantile(boots, prob=c(0.025,0.975)) ; results[3]<-CI[1] ; results[4]<-CI[2]
  if (Hist)   {  hist(boots, main="", xlab = "Значение индекса", col="grey", prob=TRUE)
  abline(v= div, col = "red",lwd=2); abline(v= div+bias, col = "red",lwd=2,lty=2)
  abline(v= CI[1], col = "green",lwd=2) ; abline(v= CI[2], col = "green",lwd=2) }
  return (results)
}
#  Получение результатов
load (file="Сок_Байт.RData") #  Загрузка численности видов по участкам из двоичного файла
# Убираем наименования видов и определяем число местообитаний
TT ; div3 <- as.matrix(TT[,-1])  ; k = dim(div3)[2]
ici("shannon", div3[,1], Hist=TRUE) ; ici("simpson", div3[,1], Hist=TRUE)
ici("shannon",method = 2,div3[,1], Hist=TRUE)
# Диаграмма доверительных интервалов для трех участков
### Таблица результатов ("пустышка")
cdiv.results <-data.frame(rbind(Diversity = rep(NA, k),Bias = rep(NA, k),
low.Limit = rep(NA, k),upp.Limit = rep(NA, k))) ; names(cdiv.results) <- colnames(div3)
### Заполнение таблицы
for (i in 1:3) cdiv.results[,i] <- ici("shannon", div3[,i], method = 2)
cdiv.results ; pldat<-data.frame(t(cdiv.results)) ;  require(Hmisc)
errbar(x = c(1:3), y = pldat$Diversity+pldat$Bias, yplus = pldat$upp.Limit,
             yminus = pldat$low.Limit, ylab = "Индекс Шеннона",
             pch=15, xaxt="n", xlab="", xlim = c(0.5, 3.5))
axis(1,c(1:3),labels=row.names(pldat))
#  Оценка зависимости точности вычисления бутстреп-оценок от числа репликаций
TB <- div3[,1]  ;  TB <- TB[TB>0] ; n=length(TB)  # Берем только Байтуган
data.b<-t(replicate(20,
     sapply(
       lapply(c(100,300,1000,5000,10000),
#  Вычисляется матрица 20x5  бутстреп-оценок индекса Шеннона
function(i) replicate(i,indiv("shannon", sample(TB, replace=TRUE)))),
                     function(i) mean(i)
#  Для получения стандартных ошибок, нижнего и верхнего доверительного интервала
#  в предыдущей строке используются функции
  #                 function(i)sqrt(var(i))
  #                function(i) sort(i)[round(length(i)*0.025)])
  #                function(i) sort(i)[round(length(i)*0.975)])
                    )))
#  График плотности вероятности бутстреп-распределения методом ядерного сглаживания
plot(density(data.b[,1]), ylim=c(0,160),col="black", lwd=2)
lines(density(data.b[,3]),col="blue", lwd=2)
lines(density(data.b[,5]),col="red", lwd=2)
#  График "ящика с усами" для разностей бутстрепируемой статистики
boxplot(sapply(1:5, function (i) (abs(outer(data.b[,i], data.b[,i], FUN="-")
                   [upper.tri(outer(data.b[,i], data.b[,i], FUN="-"))]))))
# ------------------
```

```
# ----  АВС-метод
#  Загрузка данных (численности и биомассы видов) из текстового файла с разделителями
TB <- read.table("ABC_13.txt",header=T,row.names=1,sep="\t")
library(forams) # Загрузка пакета forams с функцией abc() для расчета W-статистики Кларка
# Определение функции, выполняющей формирование бутстреп-выборки W-статистики
#  Параметры: x - объект класса abc; permutations - заданное число итераций бутстрепа
wci <- function(x, permutations=5000) {
       W <- function(v) { return(round(sum(v)/ (50 * (length(v) - 1)), 4))}
       boots <- numeric(permutations)
       for (i in 1:length(boots)) {
              boots[i] <- W(sample(x$BiAi,replace=T))}
       return (BootRes(boots, W(x$BiAi) ) ) }
wci(abc(TB))  #  Получение результатов
# -------------------
#  Расчет олигохетного индекса Гуднайта-Уитни
# Загрузка из xls-файла таблицы с исходными данными
# Столбцы:  SubClass (подкласс), Genus (род), Level (численность вида)
library(xlsReadWrite)
TT <- read.xls("D://Сок.xls", sheet = 1, rowNames=TRUE)
# Определение функции, осуществляющей расчет олигохетного индекса
indoli <- function (data) {
       ind <- 100*sum(data[data$SubClass=="Oligochaeta" & data$Genus != "Nais",
              c("Level")])/sum(data$Level);  return (ind) }
# Определение функции, выполняющей формирование бутстреп-выборки индекса J
olici <- function(data , permutations)
        { boots <- numeric(permutations)
       vyb <- t(replicate(permutations, sample.int(nrow(data), replace=TRUE)))
       boots <- sapply(1:permutations, function (i) {l<-vyb[i,]; indoli(data[l,])})
  return (BootRes(boots, indoli(data))) }
olici(TT, 5000)   #  Получение результатов
```

2. ИСПОЛЬЗОВАНИЕ РАНДОМИЗАЦИИ ДЛЯ СРАВНЕНИЯ ВЫБОРОК

2.1. Проверка статистических гипотез

Статистическая проверка гипотез является вторым после статистического оценивания параметров распределения и, в то же время, важнейшим разделом математической статистики. Если в ходе эксперимента изучаются свойства объекта, то по результатам измерений можно сформулировать некоторые содержательные предположения (*научные гипотезы*) о природе наблюдаемых закономерностей**.** Корректное подтверждение или опровержение этих предположений осуществляется с использованием *статистических гипотез*, которые логически соответствуют проверяемым научным заключениям.

Статистическими гипотезами называются предположения о свойствах распределения генеральной совокупности, которые можно принять или отклонить с минимальным риском ошибки, опираясь на выборочные данные. Нулевая гипотеза H_0 утверждает, что популяционные параметры оцениваются по результатам наблюдений, которые имеют случайно флуктуирующий характер в рамках однородной генеральной совокупности бесконечного объема или представлены в произвольном порядке. Остальные гипотезы, отличающиеся от H_0 и противопоставляемые ей, называются альтернативными и обозначаются H_1.

Выделяют (Айвазян, Мхитарян, 1998) следующие основные типы гипотез, проверяемых в ходе статистического анализа и моделирования:

1. *Об однородности* двух или нескольких обрабатываемых выборок или характеристик анализируемых совокупностей. Гипотезы однородности относительно теоретических характеристик (таких, например, как средние μ_j или дисперсии σ_j) вероятностного закона, которому подчиняются *j*-е группы выборочных наблюдений, можно записать в виде $H_{0\mu}$: $\mu_1 = \mu_2 = \ldots = \mu_j = \ldots$ или $H_{0\sigma}$: $\sigma_1 = \sigma_2 = \ldots = \sigma_j = \ldots$ Их отклонение расценивается как правдоподобное утверждение о статистической значимости группировочного фактора.

2. *О типе закона распределения* случайной величины H: $F_\xi(X) \cong F_{\text{мод}}(X, \theta)$, где $F_\xi(X)$ – исследуемая функция распределения, $F_{\text{мод}}(X, \theta)$ – модельная функция, принадлежащая к некоторому параметрическому семейству; θ – *k*-мерный параметр распределения, неизвестные значения которого оцениваются по выборке.

3. *О числовых значениях параметров* исследуемой генеральной совокупности H: $\theta \in \Delta$, где θ – некоторый одномерный или многомерный параметр, от которого зависит исследуемое распределение, Δ – область его конкретных гипотетических значений, которая может состоять из одной точки. Например, H_0: $\rho \in 0$ соответствует гипотезе об отсутствии корреляционной между двумя случайными величинами.

4. *О параметрах модели* β, описывающей статистическую зависимость между признаками. Например, отклонение гипотезы H_0: $\beta_j = 0$ может привести исследователя к предположению, что влияние рассматриваемого *j*–го фактора статистически значимо и это может появиться в данных в виде некоторой тенденции их изменчивости. Гипотезы об общем виде моделей позволяют оценить их сравнительную адекватность при выборе подходящей функции регрессии или распознавания.

Возможные исходы, которые могут иметь место при тестировании нулевой гипотезы H_0, можно представить следующим образом:

	H_0 верна	H_0 неверна
H_0 принимается	H_0 верно принята	Ошибка второго рода
H_0 отвергается	Ошибка первого рода	H_0 верно отвергнута

Часто говорят об отклонении H_0 в пользу альтернативы H_1. Акцент на альтернативную гипотезу будет зависеть от того, какие книги вы будете читать. Это понятие является центральным к школе Неймана-Пирсона, рассматривающей частотные

(frequentist) статистики. С другой стороны, Р.А. Фишер, который первым рассмотрел смысл критериев значимости в формально-систематическом ключе, никогда не оперировал альтернативными гипотезами, полностью сосредоточившись на проверке H_0. Более того, он полагал, что при проверке статистических гипотез «нулевая гипотеза сама по себе никогда не может быть принята или подтверждена, но лишь, возможно, ее удается опровергнуть» (Fisher, 1966, p. 17).

Таким образом, консерватизм научного метода проверки гипотез заключается в том, что при анализе данных мы можем сделать лишь одно правомочное заключение: нулевая гипотеза отклоняется на выбранном уровне значимости. Это не означает, что верна альтернатива H_1 – просто мы получили косвенное свидетельство ее правдоподобия на основании типичного "доказательства от противного". В случае, когда верна H_0, исследователю также предписывается сделать лишь осторожное заключение: на основе данных, полученных в условиях эксперимента, не удалось обнаружить достаточно доказательств, чтобы отклонить нулевую гипотезу. Мы по возможности в своем изложении постараемся быть терминологически политкорректными, но практический смысл не принимать на веру нулевую гипотезу, подтвержденную статистическим анализом, нам кажется излишне категоричным.

Величина, с помощью которой на основе результатов наблюдений принимается решение об отклонении нулевой гипотезы, называется *статистическим критерием*. В качестве его используют случайную величину T, являющуюся некоторой функцией данных x, плотность распределения $p_T(x)$ которой известна при справедливости H_0.

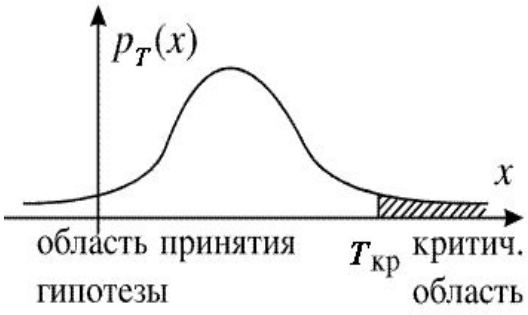


Рис. 2.1. Плотность распределения статистического критерия T при справедливости нулевой гипотезы

Пусть для некоторого априори заданного критического уровня значимости $\alpha_к$ находится критическое значение $T_{кр}$ критерия и выделяется правосторонняя критическая область. На рис. 2.1 это заштрихованная фигура, лежащая справа от точки $T_{кр}$, которая имеет площадь, равную $\alpha_к$. Если значение t_{obs}, определенное по выборочным данным, оказывается меньше, чем $T_{кр}$, то гипотеза H_0 принимается с *уровнем значимости* α, а в противном случае отвергается в пользу альтернативы. К этому правилу следует добавить два уточнения.

Различают проверку гипотез при правосторонней $Pr(T > T_{кр}) = \alpha$, левосторонней $Pr(T < T_{кр}) = \alpha$ и двусторонней $Pr\{(T > T_{кр1}) \cup P(T < T_{кр2})\} = \alpha$ альтернативах. Приведем, однако, фрагмент из дискуссии на форуме врачей-аспирантов: «…почти все статистическое тестирование рассчитано на то, что мы пытаемся ответить на вопрос: *могли ли получиться подобные различия, если бы мы брали выборки из одной популяции*. А вот односторонний критерий добавляет: *но мы учитываем при этом только положительные отклонения*… Соответственно, односторонний критерий просто игнорирует часть возможных (в рамках нулевой гипотезы) выборок. Про односторонние критерии обычно вспоминают только тогда, когда заветную границу $p = 0.05$ перейти не удалось, а очень хочется…». Применение односторонних критериев является обязательным тогда и только тогда, когда оно диктуется исключительно самой структурой анализируемых данных и соответствующих им статистических моделей (Khromov-Borisov, Henriques, 1998).

В компьютерную эпоху механизм проверки гипотез сводится к обратной процедуре: по величине t_{obs}, соответствующей выборочным данным, рассчитывается *p-значение* (*p*-value от англ. Probability – вероятность). *P*-значение – это уровень значимости α, который наблюдался бы в том случае, если бы критическое значение было выбрано равным текущему значению статистики T_{obs} (Цыплаков, 2008). Отсюда синоним – предельный уровень значимости (marginal significance level).

Другое определение *p*-значения – это «*условная вероятность* получить наблюдаемое значение t_{obs} статистики выбранного критерия *T* (и все остальные ещё менее вероятные значения этой статистики) при условии, что верна нулевая гипотеза H_0:

$$P_{val} = \Pr[|T| \geq |t_{obs.}| \,|\, H_0]\text{»}$$ (Хромов-Борисов, 2011).

Например, если необходимо оценить, насколько значимо различие средних в двух выборках, полученных из нормальных генеральных совокупностей с одинаковыми дисперсиями, рекомендуется вычислить *z*-критерий или *t*-статистику Стьюдента и определить соответствующее им значение условной вероятности *p*. Если ее величина меньше, предположим, одной тысячной, то нет веских оснований предполагать, что выборки взяты из одной генеральной совокупности. В случае, если величина вероятности *p* больше 0.05, то нельзя утверждать без серьезного риска ошибиться, что обе выборки отличаются между собой.

Здесь критическое значение, равное 5%, не является "волшебным" или научно обоснованным числом: просто люди *договорились* считать малым то, что меньше или равно 5%. В реальных условиях иногда 5%-ный уровень значимости слишком строг, а иногда слишком либерален, поэтому он должен корректироваться с учетом конкретной экспертной оценки меры ответственности за формулируемый вывод (Цейтлин, 2007).

Можно отметить несимметричность задачи проверки гипотез. Вероятность ошибки первого рода жёстко ограничивается достаточно малой наперёд заданной величиной – $Pr(p(\boldsymbol{x}) \leq \alpha \,|H_0) \leq \alpha$. Вероятность ошибки второго рода можно лишь минимизировать путём выбора достаточно мощного критерия, что часто носит субъективный характер.

Параметрические тесты, использующие традиционные статистические критерии (*t*, *z*, *F* и проч.), оценивают не то, насколько близки сами по себе данные в сопоставляемых вариационных рядах, а равны ли их отдельные выборочные характеристики. Например, если нужно сравнить две группы наблюдений при разных уровнях воздействия изучаемого фактора, то оценка отличий выборок фактически сводится к сравнению их средних (что не вполне одно и то же): т.е. формулируется гипотеза H_0: $\mu_1 = \mu_2$ и с помощью *t*-критерия делается частное заключение о равенстве центров распределения обеих групп. При использовании общепринятых непараметрических тестов (например, на основе критерия Манна-Уитни-Вилкоксона) анализ становится еще менее определенным и оперирует уже не со средними, а с таким трудно интерпретируемым и не вполне точным понятием, как "сдвиг местоположения". Важно отметить также, что после того, как рассчитан выборочный критерий t_{obs}, исходная совокупность отстраняется от дальнейшей обработки и в оценке самого *p*-значения никакого участия не принимает.

Для того чтобы корректно применять параметрические критерии, необходимо задаться целым рядом предположений: например, что обе сравниваемые совокупности распределены по нормальному закону и у них одинаковая дисперсия. Только в этих условиях *t*-статистика имеет характерное стандартное распределение в условиях справедливости нулевой гипотезы, которое вырождается (т.е. уходит в область низких вероятностей), если эмпирические данные не соответствуют H_0. Приходится либо принимать на веру нормальность и гомоскедастичность выборок, либо проверять эти утверждения с использованием других статистических критериев.

2.2. Использование метода рандомизации для проверки гипотез

Статистические тесты, разработанные Фишером (1935 г.), обеспечивают целостный и весьма здравый подход для оценки вероятности соответствия наблюдаемого объекта нулевой гипотезе. Однако, многие исследователи, в частности, Э. Эджингтон (Edgington, 1995), обосновавший технологию повторяемого случайного переприсваивания (random assignment), указывают на то, что при эксперименте в естественной среде очень редко удается получать действительно случайные выборки из генеральной совокупности. Использование параметрических критериев становится тогда теоретически не вполне корректным и они могут представлять собой лишь некоторый ориентир для более приемлемых в этих условиях методов рандомизации.

Рис. 2.2 иллюстрирует смысл случайных перестановок на примере сравнения двух малых выборок. Верхний блок показывает результат эксперимента с четырьмя значениями в основной группе и двумя значениями в контрольной. Если элементы этих групп хаотически перемешивать между собой, то можно получить два варианта перевыборок, в которых разность групповых средних меньше, чем полученная эмпирически, и один вариант – с превышением этой величины.

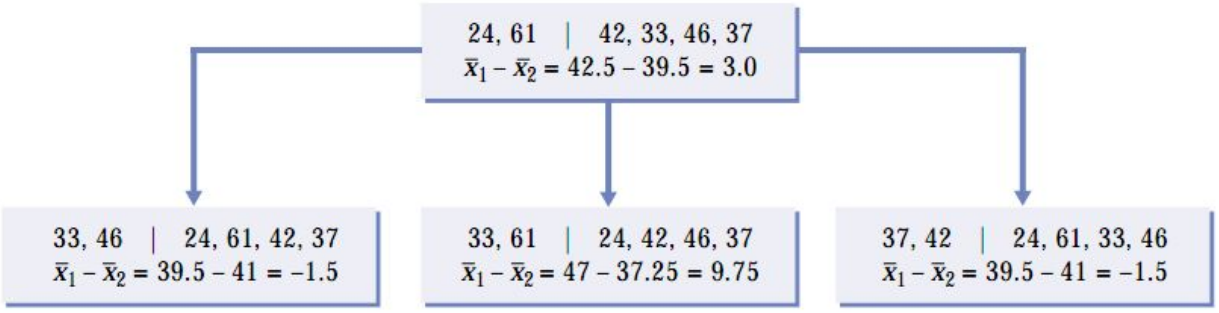


Рис. 2.2. Идея перестановочного теста. Верхний блок включает эмпирическую выборку из 6 значений, разделенных на две группы. Три нижних блока содержит псевдовыборки с теми же размерами групп, случайно составленные из исходных элементов. Рассчитана статистическая характеристика – разность выборочных средних в этих группах.

Случайное перемешивание данных между группами можно повторить многократно и получить частотное распределение критерия T (например, разности групповых средних $t = \overline{X}_1 - \overline{X}_2$). Это распределение будет иметь место при справедливости нулевой гипотезы H_0: $\mu_1 = \mu_2$, поскольку все возможные комбинации элементов в группах будут равновероятны. Тогда, если величина t_{obs} для экспериментальных данных является типичным значением из распределения рандомизации $T^*(x)$, то нет оснований отклонять H_0. Если же это не имеет место и t_{obs} оказывается в критической области критерия, то нулевая гипотеза отвергается на некотором уровне значимости и (косвенно) альтернативную гипотезу считают более разумной.

Здесь уровень значимости для t_{obs} – процент от значений из распределения рандомизации $T^*(x)$, которые являются столь же или более экстремальны, чем t_{obs}. Это соотношение может интерпретироваться таким же образом, как и для обычных критериев достоверности: при 5% превышений t_{obs} обеспечивается некоторое достаточно весомое предположение, что нулевая гипотеза не верна, а если процент превышений становится меньше 1%, то значимость отклонения H_0 становится ещё более убедительной. Мы не принадлежим «к группе людей, цель жизни которых в том, чтобы убедить остальных в неправильности 5%-го порога» (Kempthorne, Doerfler, 1969), и считаем, что лучше расценивать уровень значимости просто как подходящую меру мощности свидетельств против нулевой гипотезы, а не объявлять окончательный вердикт о достоверности влияния фактора группировки.

По сравнению с параметрическими статистическими методами у рандомизационного теста есть два главных преимущества:

- он может быть корректен без строгих требований к характеру распределения анализируемых выборок и
- при его реализации можно не ограничиваться обычными тестовыми статистиками, распределение которых может быть описано явно, а использовать для расчета критерия любую математическую формулу, отвечающую особенностям эксперимента и природе поставленной задачи.

Таким образом, рандомизация реализует процесс Монте-Карло, позволяющий получить распределение статистики анализируемого критерия, исходя из предположения,

что H_0 верна. Алгоритм выполнения рандомизационного теста для схемы сравнения двух независимых выборок объемом n_1 и n_2 состоит из следующих действий:

1. Выбираем произвольную статистику T, позволяющую оценить значимость различий выборочных средних $\overline{X}_1$ и $\overline{X}_2$ двух групп данных со стандартным отклонением $S_{\overline{X}}$ (для определенности будем использовать традиционную статистику Стьюдента $t = (\overline{X}_1 - \overline{X}_2) / S_{\overline{X}}$).
2. Вычисляем эмпирическое значение критерия для исходных сравниваемых выборок, которое обозначим как t_{obs}.
3. Повторяем B раз следующие действия, где B – число, большее, чем 1000:
   - объединяем данные из обеих выборок и перемешиваем их случайным образом;
   - первые n_1 наблюдений назначаем в первую группу, а остальные n_2 наблюдений отправляем во вторую;
   - вычисляем тестовую статистику t_{ran} для рандомизированных данных;
   - если $|t_{ran}| > |t_{obs}|$, увеличиваем на 1 счетчик b (т.е. используем двусторонний тест).
4. Разделив значение b на B, получим относительную частоту, с которой величина t_{ran} на рандомизированных данных превышает значение t_{obs} на данных, которые мы получили в эксперименте. Это соответствует оценке вероятности p того, что случайная величина $\boldsymbol{T}$ примет значение, большее, чем t_{obs}. По традиции, если $p > 0.05$, то принимается нулевая гипотеза H_0: $\mu_1 = \mu_2$ о равенстве групповых средних при использовании критерия $\boldsymbol{T}$, а если p меньше задаваемого уровня значимости, то H_0 отклоняется в пользу альтернативы.

Очевидно, что в ходе случайных переприсваиваний не меняется ни состав исходных выборок, ни численность групп с разными уровнями воздействия, а только происходит беспорядочный обмен элементами данных между этими группами. Интересно, что с равным результатом вместо традиционного t-критерия (при равенстве дисперсий в группах) можно использовать простую разность между средними $(\overline{X}_1 - \overline{X}_2)$.

Часто для описанного алгоритма рандомизации употребляют термин *перестановочный* или пермутационный (permutation) тест, имея в виду перестановку данных между отдельными группами. Этот термин не вполне удачен, поскольку в действительности мы осуществляем не перестановки, а берем различные *комбинации* данных, уникальных относительно выбранной тестовой статистики. В частности, в примере на рис. 2.2 перестановка {24, 61} и {61, 24} в одной из групп не является шагом рандомизации, поскольку приводит к тем же значениям групповых средних, медиан, вариаций и т.д. Фраза "рандомизационный тест", как отмечает Д. Ховел, является хорошим компромиссом, чтобы избегнуть двусмысленности термина "перестановка".

В ходе рандомизационного теста вычисляется в некотором смысле "точное" значение уровня значимости p, непосредственно определяемое анализируемыми выборками. Если мы имеем множество из $B = 99$ рандомизированных значений тестовой статистики, то уровень значимости H_0, вычисляемый по скорректированной формуле

$$p = (b + 1)/(B + 1) \qquad \text{(Davison, Hinkley, 1997)},$$

соответствует количеству элементов этого множества b, равных или превышающих эмпирическое значение. Добавление 1 в числитель и знаменатель позволяет не только учесть на равных правах исходную выборку, но и избежать появление нулевых вероятностей (число итераций B принято в этом случае назначать как цены в супермаркете 999, 4999, 9999 и т.д.). Можно также отметить, что расчет вероятности ошибки 1 рода по формуле Дэвисона-Хинкли как нельзя лучше соответствует первоначальным взглядам Фишера (Fisher, 1926) на p-значение, просто как на неформальный индекс, оценивающий меру согласия эмпирических данных с нулевой гипотезой.

В случае небольших выборок можно использовать полный перебор всех возможных комбинаций данных, вычисляя для каждой из них тестовую статистику, что,

разумеется, часто оказывается невозможным. Например, если у нас есть три группы с 20 наблюдениями в каждой, то мы имеем 60!/(20!·20!·20!) или $5.78 \cdot 10^{26}$ различных вариантов группировки наблюдений, и даже быстрый суперкомпьютер будет не в состоянии их перебрать. Решение состоит в том, что мы берем ограниченную случайную выборку из всех возможных комбинаций, которая не будет приводить к *точному* ответу. В этом случае рандомизацию, как и непараметрический бутстреп, можно трактовать как разновидность имитационного процесса Монте Карло.

Однако сколько перевыборок мы должны выполнить, чтобы гарантированно оценить уровень значимости достаточно близко к его истинному значению для анализируемых выборок? Эджингтон (1995, p. 55) показал, что оценка уровня значимости p рандомизационного теста будет распределена приблизительно по нормальному закону с дисперсией $p\ (1 - p)/B$, если число итераций B достаточно велико. Следовательно, 99% оценок уровней значимости будет в интервале $p \pm 2.58\sqrt{p(1-p)/B}$, на основании чего легко предположить, что $B = 5000$ является разумным минимумом для испытания на 1%-ом уровне. Однако результаты уже 1000 итераций могут удовлетворить не слишком придирчивого исследователя, поскольку это – разумный минимум для рандомизации на 5%-ом уровне значимости, а погрешность тестовой статистики будет наблюдаться лишь в 3-м десятичном разряде или менее того.

Отметим однако, что доля итераций, в которых случайная величина T оказывается столь же или более экстремальна чем t_{obs}, не в полной мере соответствует смыслу проверки гипотезы о различии между внутригрупповыми средними, т.к. статистика t здесь уже используется просто как один из подходящих индексов, измеряющих некую обобщенную "неодинаковость" выборок. Точно так же можно оценить различие в группах с использованием выборочных медиан, дисперсий или коэффициентов вариаций, любых метрик сходства выборок и проч. «При рандомизацонном тесте нулевая гипотеза выражается более свободно. Я формулирую ее просто: ‘группировка не влияет на значения наблюдаемой переменной’, не уточняя, имеется ли при этом в виду среднее, медиана, дисперсия или даже форма распределения. Я оставляю это в значительной степени в воздухе.» – так обосновывает эту точку зрения Д. Ховел.

Рандомизация рассматривает понятие нулевой гипотезы шире, чем простая проверка предположений относительно конкретных параметров распределений. Сделаем предварительно небольшой философский экскурс и напомним следующие концептуальные основы применения нуль-моделей (Gotelli, Graves, 1997):

° *Нуль-модель* – метод формализации и последующей проверки нулевой гипотезы, утверждающей, что в системе не произошло никаких изменений, либо эти изменения нельзя приписать влиянию рассматриваемого фактора.

° Нуль-модель каждый раз проектируется определенно заданным образом, чтобы скомпенсировать потенциальное воздействие конкретных изучаемых процессов или предполагаемых причин. Для этого выполняется случайное перемешивание исходных данных или формируется рандомизированная выборка из некоторого распределения.

° В то же время, нуль-модель – это некий сценарий формы существования или развития системы в условиях делокализации изучаемых факторов, принимая во внимание, что все остальные механизмы и внутрисистемные связи не являются нарушенными.

° Если математические модели выдвигают на первый план влияние определенных факторов или процессов и явно включают в себя задачу оценки их параметров, то нуль-модели, напротив, преднамеренно исключают эти механизмы, чтобы оценить получившийся результат. Математическим моделям иногда не требуется никакой эмпирической информации для их разработки, тогда как нуль-модели создаются всегда в отношении конкретного набора данных.

° Нуль-модели, ориентированные на отсутствие эффекта воздействия, всегда полагаются на принципы экономности продуктивных гипотез и их непременной

фальсифицируемости и настойчиво подчеркивают потенциальную значимость стохастических механизмов в функционировании естественных систем. Подробно характеристики нуль-моделей и смысл описываемых ими гипотез представлены в работах (Wright et al., 1998; Gotelli, 2000; Gotelli, Entsminger, 2003; Miklós, Podani, 2004).

Таким образом, процедура рандомизации в общем виде состоит из трех шагов:

- выбор выражения для критериальной статистики T;
- разработка способа имитации структуры наблюдаемых данных, т.е. нуль-модели, адекватной поставленной задаче и сформированный исходя из предположения, что H_0 верна (это может быть не только простая перестановка выборочных значений, но и достаточно серьезные алгоритмы – см. раздел 2.8);
- запуск процесса Монте-Карло, позволяющего по серии из B реализаций нуль-модели восстановить функцию плотности распределения анализируемого критерия $p_T(x)$.

2.3. Сравнение статистических характеристик двух независимых выборок

Взятие гидробиологических проб часто осуществляется из предположения, что на различных участках рек видовое богатство гидробионтов может меняться. Фактически это – разумная гипотеза, вытекающая из неоднородности условий среды обитания и концепции речного континуума. Однако требуется статистический анализ, насколько нулевая гипотеза, утверждающая, что различия видовой структуры носят случайный характер, совместима с данными гидробиологической съемки и есть ли веские причины отклонить ее в пользу альтернативной гипотезы, что эти различия существуют.

Предположим, что необходимо выяснить, отличается ли точечное видовое разнообразие донных организмов для верхнего (51 проба) и нижнего (44 пробы) участков р. Сок [пример П2]. Сформируем две выборки из значений индекса Шеннона для каждой пробы зообентоса и рассчитаем для обоих участков средние величины индексов разнообразия $\overline{X}_1 = 2.251$, $\overline{X}_2 = 2.475$, а также их выборочные стандартные отклонения $S_1 = 0.8$ и $S_2 = 0.936$.

Предварительно проверим различие видового разнообразия между участками обычными параметрическими методами. Предположим, что индексы Шеннона представляют собой случайные выборки из нормальных распределений $N(\mu_1, \sigma^2_1)$ и $N(\mu_2, \sigma^2_2)$ соответственно. Тогда интересующий нас комплект гипотез оценивает равенство средних: H_0: $\mu_1 = \mu_2$ против альтернативы H_1: $\mu_1 <> \mu_2$. Нулевая гипотеза может быть проверена с использованием критерия Стьюдента с поправкой Уэлча для неравных дисперсий:

$$t = \frac{(\overline{X}_1 - \overline{X}_2)}{\sqrt{(n-1)n_1 S_1^2 + (n_2 - 1)n_2 S_2^2}} \cdot \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}} = 1.253.$$

Если H_0 верна, то статистика критерия является случайным значением из t-распределения с $(n_1 + n_2 - 2)$ степенями свободы. Вероятность того, что наблюдаемое значение будет превышено, составляет $p = 0.219$, и поэтому нет оснований для отклонения нулевой гипотезы.

Другой вариант параметрического теста состоит в оценке 95%-х доверительных интервалов для разности групповых средних:

$$\theta_H = \left|\overline{X_1 - X_2}\right| - t_{\alpha,df} S_D = -0.13 \quad \text{и} \quad \theta_B = \left|\overline{X_1 - X_2}\right| + t_{\alpha,df} S_D = 0.577,$$

где число степеней свободы df и ошибка средней разности S_D рассчитываются по формулам для выборок разного объема; $\alpha = (1 - 0.95)/2$. Поскольку доверительный интервал разности средних включает 0, то нельзя говорить о ее статистической значимости.

Но напомним предположения, которые сделаны в этом анализе:

1. Случайный отбор гидробиологических проб на сравниваемых участках рек;
2. Равные стандартные отклонения обоих выборочных совокупностей;
3. Нормальный закон распределения индексов Шеннона в пределах групп.

Предположение (1) обычно сомнительно, так как сбор гидробиологических данных далеко не всегда отвечает случайному выборочному процессу. Предположение (2) может оказаться верным: по крайней мере, дисперсионное отношение Фишера $F = S_1/S_2$. = 0.936/0.8 = 1.366 статистически значимо не отличается от 1 (p = 0.287). Предположение (3) может быть приблизительно верным, но это не может быть серьезно проверено на основе имеющихся небольших выборок. В целом представленные тесты могут оказаться весьма устойчивыми к перечисленным предположениям, но иногда возможные отклонения в выводах могут оказаться фатальными, особенно когда объемы выборок являются неравными.

Рандомизационный тест основан на операции случайного переприсваивания (random sampling without replacement), которая состоит в том, что исходные данные обеих выборок объединяются, все значений 95 индекса Шеннона хаотически перемешиваются, после чего снова распределяются по двум группам в том же соотношении n_1 к n_2., равном 1.25. Этот процесс повторяется многократно, формируется 5000 пар псевдовыборок, и каждый раз вычисляется нормированная разность средних между этими группами (т.е. t-статистика).

В результате получаем (рис. 2.3) частотное распределение t-статистики для рассматриваемого блока данных при справедливости нулевой гипотезы, поскольку случайное переприсваивание имитировало условия, при которых видовое разнообразие однородно на всех участках реки. Экспериментально найденное значение t-критерия расположено где-то внутри этого распределения, в частности 1089 нуль-модельных комбинаций из 5000 превышают по абсолютной величине эмпирическую величину. Следовательно p-значение равно 0.2178, отклонять нулевую гипотезу нельзя и мы делаем вывод, что среднее видовое разнообразие зообентоса в верхнем и нижнем течении р. Сок не отличается между собой. Еще раз подчеркнем, что в случае рандомизационного теста нам нет необходимости проверять предположения о нормальности распределения выборок и равенстве их дисперсий (Everitt, Howell, 2005, pp. 1674-1681).

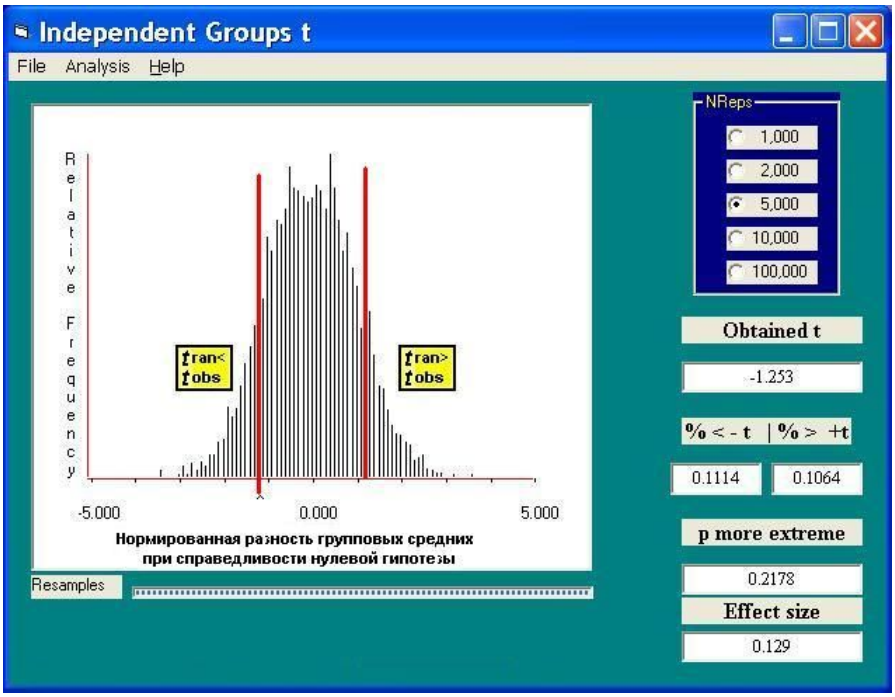


Рис. 2.3. Распределение t-статистики, оценивающей разность групповых средних индекса Шеннона при справедливости нулевой гипотезы. Красными линиями показано положение наблюдаемой t-статистики (Obtained t); P more extreme – вероятность превышения наблюдаемой статистики; Effect size - доля эффекта от группировки данных

Аналогичные расчеты для другой реки Чапаевка показывают, что видовое разнообразие в верхнем течении значительно превышает этот показатель в нижнем течении. В результате 5000 случайных перестановок 244 значений индекса Шеннона между двумя группами, не нашлось ни одной такой комбинации, чтобы различия оказались бы больше, чем в реальных данных, а эмпирическая *t*-статистика расположилась далеко за пределами нуль-модельного распределения.

Отметим, что в обоих примерах *p*-значения, полученные параметрическим и рандомизационным тестом оказались чрезвычайно близки между собой, что свидетельствует о стабильности и корректности как того, так и другого метода. Такая близость отмечается в большинстве случаев, особено если тестируемые данные отвечают требованиям "параметризуемости". Однако если выборки наблюдений содержат одно или несколько аномальных значений, тогда, возможно, итоговые выводы окажутся не столь согласованными. Эджингтон (1995, p. 11) считал единственными правомочными классическими методами только те, которые дают те же результаты, что и испытания рандомизации, хотя это может показаться довольно экстремальной точкой зрения.

В анализе по р. Чапаевка коррекция Дэвисона-Хинкли для выражения $p = (b+1)/(B+1) = 1/5000 = 0.0002$ позволяет избежать появления нулевой вероятности. В случае параметрических критериев *p*-значение находится косвенными методами по приближенным формулам аппроксимации: например, в том же примере значению $t = 8.95$ при 242 степенях свободы соответствует вероятность $p = 9 \cdot 10^{-17}$.

Если в случае традиционных методов проверки гипотез обычно конкретизируется вид полуэвристических выражений для статистики критерия и последующий способ оценки *p*-значения (с помощью подходящего теоретического распределения или тех же полуэвристических формул), то перестановочный тест, не имеющий этих логических нагромождений, с одинаковой легкостью может использовать любую подходящую формулу. В приведенных примерах мы использовали *t*-статистику лишь для возможности сравнить разные методы, но точно такие же *p*-значения можно получить, если использовать простую разность средних или даже среднее для первой выборки. Применение последнего даже предпочтительнее, поскольку избавит компьютер от ненужной вычислительной работы.

Весьма наглядные результаты могут получиться при использовании в тесте отношения групповых средних $\overline{X}_1 / \overline{X}_2$, которое показывает, насколько центральная тенденция в первой выборке превышает эту величину во второй. В случае индекса Шеннона для участков р.Сок наблюдаемое отношение 2.251/2.475 = 0.91. Программа RundomPro дает возможность выполнить на этих данных 5000 итераций рандомизации и получить распределение статистики $\overline{X}_{1ran} / \overline{X}_{2ran}$ при справедливости нулевой гипотезы H_0: $\mu_1/\mu_2 = 1$ – рис. 2.4. Основываясь на этом распределении, можно заключить, что вероятность получить еще меньшее отношение, чем эмпирическое $\overline{X}_1 / \overline{X}_2 = 0.91$, составляет $p_l = 0.1059$, а еще большую величину - $p_r = 0.894$. Поскольку нельзя сказать, что это отношение статистически значимо отличается от 1, то отвергать нулевую гипотезу об однородности разнообразия макрозообентоса в р. Сок нет оснований. Поскольку полученные *p*-значения очень близки к оценкам, установленным ранее на основе использования *t*-статистики, то оба тестовых критерия можно счесть эквивалентными.

Разумеется, далеко не все выражения для тестовой статистики являются эквивалентными относительно получаемого результата и это необходимо учитывать при планировании вычислений. Например, еще одна схема сравнения двух групп может быть основана на их медианной разности. Медиана индексов Шеннона для верхнего течения р. Чапаевки $M_1 = 2.45$, для нижнего $M_2 = 1.6$, а медианная разность разнообразия на обоих участках ($M_1 - M_2$) = 0.85 (рис. 2.5).

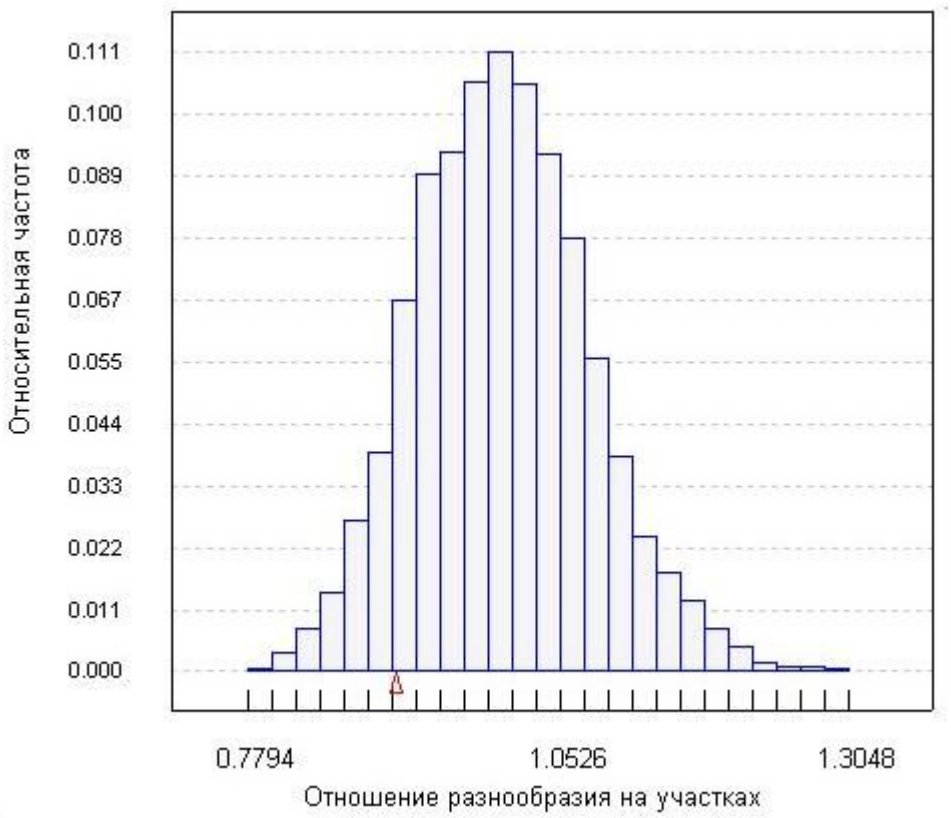


Рис. 2.4. Распределение отношения средних значений индекса Шеннона для р. Сок при справедливости нулевой гипотезы. Треугольником отмечено положение наблюдаемой статистики. (расчеты выполнены по программе RundomPro 3.14)

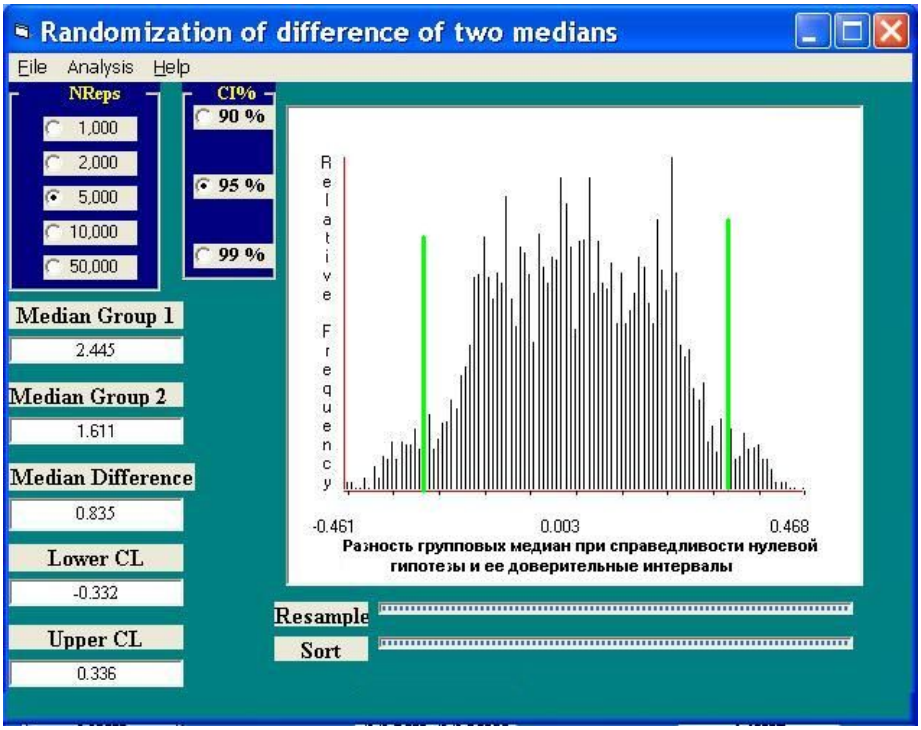


Рис. 2.5. Оценка доверительных интервалов медианной разности (Median Difference) двух выборок значений индекса Шеннона при справедливости нулевой гипотезы

Оценка статистической значимости медианных разностей может быть выполнена по тому же механизму, что и рандомизационный тест на сравнение средних. Разумеется, мы уже не можем использовать *t*-статистику, поскольку нет корректного способа вычислить стандартную ошибку. Хорошей альтернативой является восстановить

распределение медианных разностей при справедливости нулевой гипотезы H_0: $M_1 = M_2$ и подсчитать число итераций, при которых эта разность была бы столь же большой (или больше), чем в эмпирических выборках.

Однако можно пойти и другим путем: вычислить доверительные границы разности медиан и отклонить нулевую гипотезу, если эмпирическое различие окажется вне доверительного интервала. Преимущество этого подхода состоит в том, что он дает нам интервальные оценки, которые всегда полезно иметь. С другой стороны, при таком подходе можно отклонить H_0, если разность медиан фактических выборок окажется вне доверительной области, но мы не получим точного *p*-значения вероятности ошибки 1-го рода.

Граничные значения X_H^* и X_B^* доверительных интервалов могут быть вычислены описанным в главе 1 способом, который Б.Эфрон называет "методом процентилей". Если случайным образом 5000 раз переставлять значения между группами, то получим нуль-модельное распределения медианных разностей индекса Шеннона для двух участков реки. Нижнюю границу X_H^*= -0.332 двустороннего 95%-го доверительного интервала межгрупповой разности медиан можно найти, если в ранжированном ряду рандомизированных разностей отсчитать 125-ое (0.025·5000) порядковое значение, а верхнюю границу X_B^*= 0.336 – выбрав 4875-й (0.975·5000) член этого ряда. Очевидно, что эмпирическая разность медиан (0.85) выходит далеко за пределы этих доверительных интервалов, следовательно нулевая гипотеза и здесь может быть отвергнута.

Здесь следует отметить важное обстоятельство: формально доверительные границы были рассчитаны нами не для истинной разницы медиан, а в предположении, что *верна нулевая гипотеза*, и поэтому их познавательная ценность ограничена. Установить доверительные пределы *истинной медианной разности* для описанных выборок можно, например, с использованием бутстрепа: при 95%-ом уровне надежности они будут в границах от 0.604 до 1.129 (см. рис. 2.6). Поскольку этот интервал не включает 0, мы также можем отклонить H_0, но необходимо ясно представлять, сколь принципиально различны два подхода, реализующие рандомизацию и бутстреп.

В общем случае выборочную оценку параметров положения $\hat{\theta}$, $\hat{\theta} = \sum_{i=1}^{n} x_i w_i$, можно интерпретировать как расчет весов w_i для каждого *i*-го члена вариационного ряда наблюдений x_i. Веса обычно являются некоторой функцией от текущих значений x_i в вариационном ряду, обычно задаются на основе предположений о законе распределения случайной величины и подчиняются правилам нормировки $\sum_{i=1}^{n} w_i = 1$. Для нормального распределения w_i – это относительные частоты появления каждого значения. Для равномерного распределения $w_1 = w_n = 0.5$, а остальные веса равны нулю и оценка меры положения равна полусумме минимального и максимального значений. Для выборочной медианы также достаточно положить нулю все веса w_i , кроме одного ($w_{(n+1)/2} = 1$, если *n* нечетное) или двух ($w_{n/2} = w_{n/2+1} = 0.5$, если *n* четное).

Любые методы статистического анализа (рандомизационные тесты тут не являются исключением) чувствительны к возможным выбросам или иным аномальным значениям. Чтобы скомпенсировать этот эффект, проводят цензурирование (censoring) выборок, которое сводится к присвоению нулевых весов хвостовым членам вариационного ряда, тогда как остальным приписываются одинаковые положительные веса, т.е. $w(x_i) = w_0$, если $a \le x_i \le b$ и $w(x_i) = 0$, если $b \le x_i$ или $x_i \le a$. Границы выделяемого интервала [*a*, *b*] часто задают с использованием квантилей, обрезая, например, слева и справа по 25% экстремальных значений. Оценки параметров, построенных по цензурированным выборкам, хотя и не являются наилучшими в жестких рамках генеральной совокупности определенного типа, но обладают выгодными свойствами устойчивости по отношению к тем или иным отклонениям от априорных допущений.

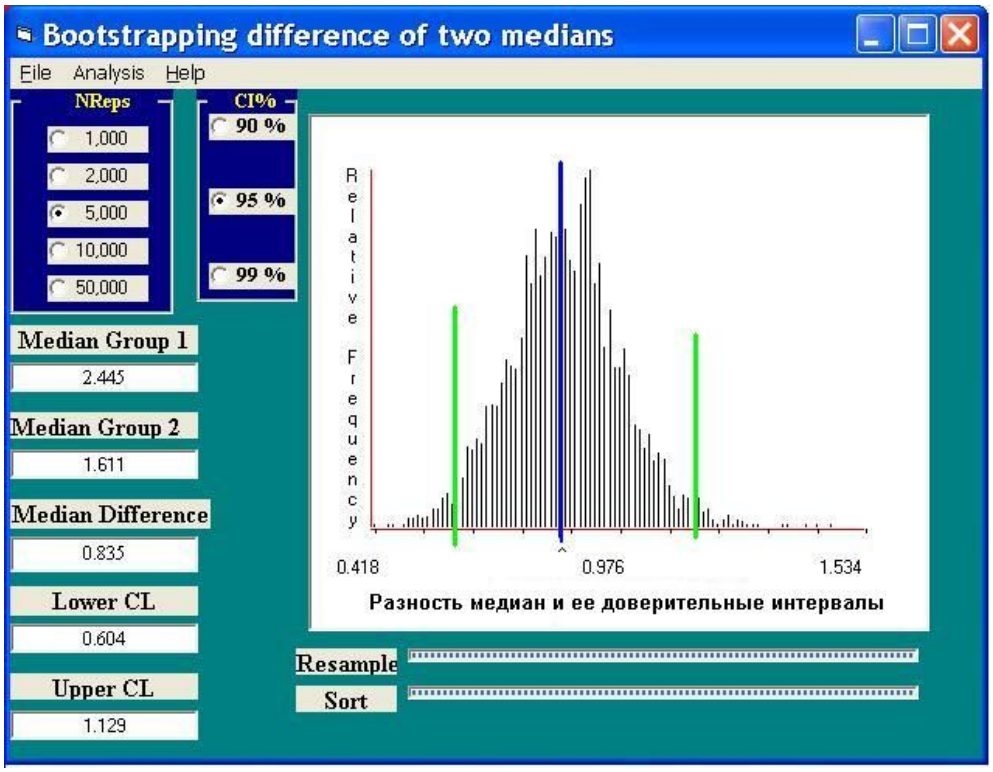


Рис. 2.6. Оценка доверительных интервалов медианной разности (Median Difference) двух выборок значений индекса Шеннона бутстреп-методом

Выполним сравнение общей биомассы зоопланктона в двух точках Куйбышевского водохранилища [пример П1]: в районе впадения р. Кама (ст. № 51, $n_1 = 61$) и у плотины Куйбышевской ГЭС (ст. № 34, $n_2=123$) по данным гидробиологического мониторинга в период 1958-1984 гг. Параллельно воспользуемся для этого схемами рандомизации, реализованными в программе RundomPro 3.14 и в кодах R к этому разделу, а результаты проверки нулевой гипотезы представим в табл. 2.1.

Таблица 2.1. Достигнутый уровень значимости при рандомизационном тесте (5000 итераций) односторонней и двухсторонней гипотез об однородности биомасс зоопланктона в двух точках Куйбышевского водохранилища

Устье Камы (№ 51)		Плотина ГЭС (№34)		Разность $d_{obs} = m_1 - m_2$	$P\ (d_{ran} > d_{obs})$	$2P\ (d_{ran} > d_{obs})$	$P\ (\|d_{ran}\| > \|d_{obs}\|)$
Объем, n_1	Среднее m_1	Объем, n_1	Среднее m_2				
Нецензурированные выборки							
61	0.882	123	0.413	0.469	0.0098	0.0196	0.0196
Выборки, усеченные справа на 5%							
58	0.538	117	0.315	0.223	0.0008	0.0016	0.0012
Выборки, усеченные симметрично по квартилям							
31	0.358	62	0.238	0.119	0.0004	0.0008	0.0004

Распределение разности выборочных средних при справедливости H_0 на рис. 2.7 имеет характерную бимодальную форму, имеющую место при наличии аномально высокого значения, которое поочередно случайно попадает то в одну, то в другую группу.

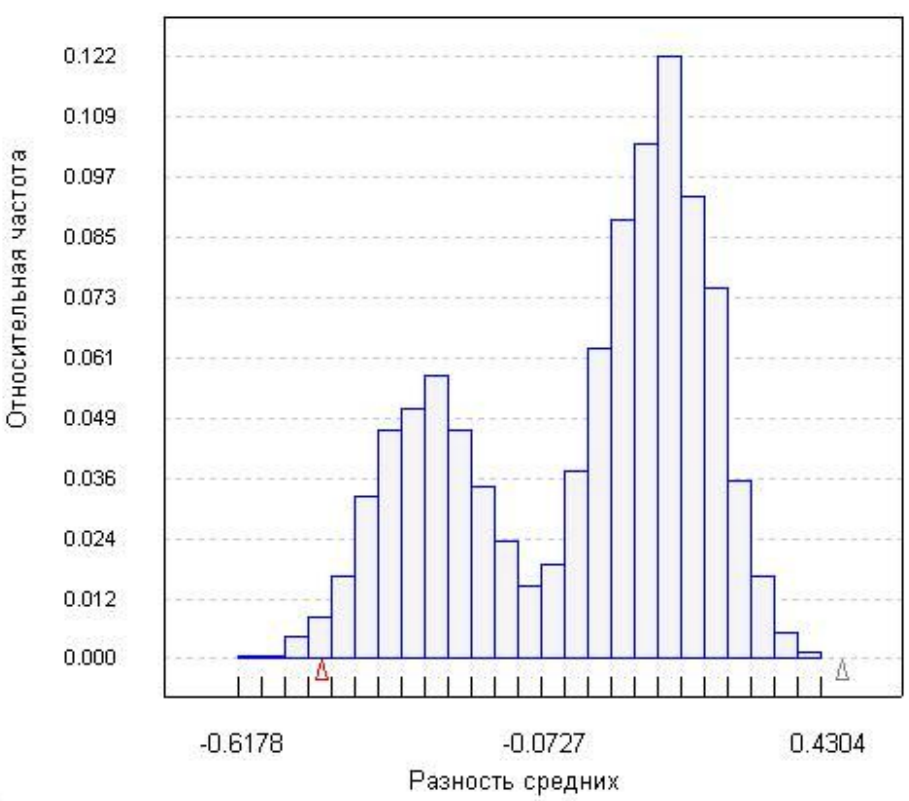


Рис. 2.7. Распределение разности групповых средних биомассы зоопланктона на станциях наблюдения №51 и №34 Куйбышевского водохранилища при справедливости нулевой гипотезы. Треугольником отмечено положение наблюдаемой статистики.

Это распределение становится симметричным и унимодальным, если выполнить одностороннее цензурирование выборок, отбросив в каждой из них по 5% максимальных значений, а именно: на станции 51 – 17.7, 2.6 и 2.3 г/м^3; на стнции № 34 - 3.1, 3, 2.3, 2.1, 1.9 и 1.6 г/м^3. При этом вывод о превышении обилия зооплактона в устье Камы над его плотностью в районе г. Тольятти приобрел дополнительную статистическую значимость: $p = 0.0008$ против $p = 0.0098$ для нецензурированных выборок. Впрочем, после еще более глубокого двухквартильного "урезания" выборок достоверность различий еще более проявилась, что наводит на мысль о некоторых "странностях" в мире статистики.

К разделу 2.3:

```
# Сравнение статистических характеристик двух независимых выборок
# Определяем два вектора индексов Шеннона для участков р. Сок
vec1 <- c(2.298,2.805,3.031,2.378,1.676,1.91,2.95,3.479,3.305,2.142,2.786,2.377,0.369,1.421,
1.877,2.192,2.218,2.035,2.932,3.114,3.059,2.385,2.506,2.716,1.696,1.829,1.342,1.601,3.463,
2.925,0.224,2.833,2.899,2.822,2.708,1.829,2.049,0.523,3.096,2.426,0.991,2.735,1.129,2.303,
1.889,2.497,0.715,3.039,2.073,1.857,3.363)
vec2 <-
c(2.865,3.183,2.752,2.866,0.918,0.946,1.19,1.585,4.026,2.583,3.605,4.035,3.88,3.74,1.807,
2.823,2.015,0.948,2.734,0.9,2.7,2.047,2.482,0.752,1.251,2.776,3.518,2.153,2.808,1.639,3.431,
1.676,2.873,2.733,3.778,2.999,1.607,2.645,2.104,3.628,2.089,2.377,3.475,1.933)
#                    Параметрический тест
#  Функция для расчета t-статистики (Версия P.Legendre, 2005)
t.stat <- function(n1,n2,vec1,vec2) {
       moy1 <- mean(vec1) ; moy2 <- mean(vec2) ; var1 <- var(vec1) ; var2 <- var(vec2)
       var.wm <- ( (n1-1)*var1 + (n2-1)*var2 ) / (n1+n2-2)
       t <- ( moy1-moy2 ) / sqrt(var.wm * ((1/n1) + (1/n2)) )
       return(list(moy1=moy1,moy2=moy2,var1=var1,var2=var2,stat=t))
}
# Вывод результатов параметрического теста
n1 <- length(vec1) ; n2 <- length(vec2) ; n  <- n1+n2 ; t.ref <- t.stat(n1,n2,vec1,vec2)
```

```
p1 <- pt(t.ref$stat,(n-2),lower.tail=TRUE) ; p2 <- pt(t.ref$stat,(n-2),lower.tail=FALSE)
p3 <- ifelse(p1 > p2, p2*2, p1*2)
c(cat("Размеры выборок:",n1,n2,"\n"), cat("Групповые средние:",t.ref$moy1,t.ref$moy2,"\n"),
         cat("Групповые дисперсии:",t.ref$var1,t.ref$var2,"\n"),
         cat("t =",t.ref$stat,"  d.f. =",(n-2),"\n"),
         cat("Односторонняя гипотеза (t слишком мало)  p = ",p1 ,"\n"),
         cat("Односторонняя гипотеза (t слишком велико)  p = ",p2,"\n"),
         cat("Двусторонняя гипотеза   p = ",p3,"\n"))
#  Те же результаты получаем с использованием стандартной функции R
t.test(vec1,vec2,var.equal=TRUE)
#  Непараметрический тест Вилкоксона-Манна-Уитни
wilcox.test(vec1,vec2, conf.int = TRUE)
#                 Рандомизационный тест
source("print_rezult.r")  # Загрузка функций вывода результатов
#  Функция, определяющие различные варианты статистик для тестирования
CompStat <- function (v1, v2, method = 1) {
   if (method == 1) st <- t.stat(length(v1),length(v2),v1,v2)$stat #  Статистика Стьюдента
   if (method == 2) st  <- sum(v1) -                                #  Разность
   if (method == 3) st <- median(v1) - median(v2)                  #  Разность медиан
   if (method == 4) st  <- mean(v1)/mean(v2)                       #  Отношение средних
  return (st) }
#  Функция, выполняющая рандомизационный тест заданной статистики заданное число раз
simP <- function(vec1, vec2, permutations=5000, method=1) {
 empar <- CompStat(vec1, vec2, method) ; boots <- numeric(permutations) ; vec <- c(vec1,vec2)
  n1 <- length(vec1) ; n2 <- length(vec2) ; n  <- n1+n2
  for (i in 1:permutations) {
#   Каждый раз заменяем значения обеих выборок случайными перестановками общего вектора
     vec.perm  <-  sample(vec,n) ;  vec1.perm <- vec.perm[1:n1] ;
     vec2.perm <- vec.perm[(n1+1):n]
     boots[i] <- CompStat(vec1.perm, vec2.perm, method) }
  return (RandRes (empar, boots, permutations)) }  #    Вывод результатов
#   Выполнение рандомизации
simP(vec1, vec2,5000,1) ; simP(vec1, vec2,5000,2)
simP(vec1, vec2,5000,3); simP(vec1, vec2,5000,4)
#   Функция цензурирования выборки
# Параметры: v  - выборка , p.left - доля, урезаемая слева, p.right - доля, урезаемая справа
censor <- function (v, p.left=0, p.right=0.05) { v.c <- v
       if (p.left>0) v.c <- v.c[which(v.c >= quantile(v, prob=p.left))]
       if (p.right>0) v.c <- v.c[which(v.c <= quantile(v, prob=1-p.right))]
       return (v.c) }
#  Расчет по выборкам, обрезанным на 5% справа
v.c1 <- censor(vec1, 0, 0.05) ; v.c2 <- censor(vec2, 0, 0.05) ; simP(v.c1, v.c2)
#  Расчет по выборкам, обрезанным на 25% с обоих сторон
v.c1 <- censor(vec1, 0.25, 0.25) ; v.c2 <- censor(vec2, 0.25, 0.25) ; simP(v.c1, v.c2)
```

2.4. Рандомизационный тест для связанных выборок

Впервые идеи рандомизации обсуждались Фишером (1935 г.) именно на примере связанных выборок. Предположим, что имеется *n* объектов, для которых значение изучаемого показателя было измерено до и после реализации некоторого воздействия, т.е. имеется *n* сопряженных пар наблюдений. В частности, Ховел приводит пример лечения анорексии с использованием когнитивной терапии поведения (Cognitive Behavior Therapy), которая может сопровождаться изменением массы тела пациентов. Другой пример связан с экспериментом Чарльза Дарвина, когда в парах растений *Zea mays* одинакового возраста и от одних и тех же родителей один экземпляр подвергался перекрестному опылению, а другой – самооплодотворению.

Сразу оговоримся, что можно разделять естественное воодушевление, что благодаря предложенной терапии легко потолстеть, питаясь лишь познанием (cognitive – познавательный), а также принять на веру вывод о том, что перекрестное опыление приводит к более крупному потомству. Однако эти примеры являются ярким образцом

некорректной экспериментальной методологии. Классическая концепция проведения эксперимента такова: любой управляемый эксперимент должен иметь повторности, причем группы экспериментальных единиц формируются случайным образом и для каждой из них также случайно должны быть назначены различные уровни изучаемого воздействия, включая обязательную контрольную группу (Hurlbert, 1984). Но в нашем первом примере все пациенты получали одно и то же лечение, контрольной группы укомплектовано не было, поэтому нет никакого основания утверждать, что увеличение массы тела произошло вследствие когнитивной терапии, а не по причине каких-то иных факторов (например, пациентов просто хорошо кормили). Во втором примере также нет оснований утверждать, что исследуемые растения были идентичны. Вряд ли можно увидеть подобные примеры в серьезной литературе, в частности, в книге Эджингтона (Edgington, 1995) по тестам рандомизации, где постоянно подчеркивается необходимость случайного назначения воздействий экспериментальным единицам.

Параметрический *t*-тест для сопряженных пар наблюдений сводится к анализу выборки, составленной из разностей: $t = \frac{\overline{d}}{s\sqrt{n}}$; $s = \sqrt{\frac{1}{n-1}\sum_{1}^{n}(d_i - \overline{d})^2}$; $d_i = x_{2i} - x_{1i}$.

Если верна нулевая гипотеза H_0: $D = 0$, утверждающая, что средняя разность D между парами реализаций случайных величин статистически значимо не отличается от нуля, то нет оснований предполагать, что эффект воздействия имеет место.

На такой же простой идее основывается и рандомизационный тест: если исследуемый фактор не имеет никакого влияния на характер данных, то с равной вероятностью величина показателя, измеренного у любого объекта *после* воздействия, будет больше или меньше значения показателя у того же объекта *до* нанесения воздействия. Другими словами, если нулевая гипотеза верна, то перестановка данных в пределах любой *пары* равновероятна и приводит к одинаковому итоговому результату.

Если зафиксировать друг с другом все пары измерений и менять местами измерения ДО и ПОСЛЕ воздействия в одной или нескольких случайно выбранных парах, вычисляя каждый раз значение тестовой статистики t_{ran}, то после многократных перестановок можно восстановить ее нуль-модельное вероятностное распределение (другое название – reference distribution). На основе этого распределения оценивается, какую вероятность составляет получение величины той же статистики t_{obs} для эмпирически измеренных данных.

Рассмотрим в качестве примера изменение численности (млн. клеток/л) сине-зеленых водорослей в составе фитопланктона Куйбышевского водохранилища [пример П1]. По данным отбора гидробиологических проб в течение июля-августа рассчитаем среднее обилие организмов на каждой из 17 станций наблюдения за периоды 1974-1979 гг. (выборка 1) и 1980-1984 гг. (выборка 2), т.е. до и после ввода в строй Чистопольской и Новочебоксарской ГЭС – см. табл. 2.2.

Проверка гипотезы, выполненная с использованием критерия Стьюдента, дала $t_{obs} = 1.576$, что после аппроксимации теоретическим распределением соответствует $p = 0.135$. Во многих руководствах по статистической обработке рекомендуется параллельно провести анализ на основе непараметрических критериев: критерия знаков ($r_{\pm} = 11$; $p = 0.332$) или статистики Вилкоксона-Манна-Уитни ($W = 106$; $p = 0.163$).

Выполнение 5000 циклов рандомизации дает возможность построить гистограмму выборочного распределения *t*-статистики при справедливости H_0 (рис. 2.8) и получить достигнутый уровень значимости $p = 0.155$, несколько более высокий, чем при использовании асимптотики. В случае односторонней гипотезы H_0, что обилие сине-зеленых после 1980 г. не будут превышать зарегистрированные значения в предыдущем периоде, риск ошибиться составит 416/5000 = 0.0832, где 416 – число итераций рандомизации, в которых имитируемая *t*-статистика превысила эмпирическое значение.

Таблица 2.2. Численность сине-зеленых водорослей Куйбышевского водохранилища в разные периоды наблюдений

Станция	X_1 1974-79 г	X_2 1980-84 г	Разность $D > 0$	Станция	X_1 1974-79 г	X_2 1980-84 г	Разность $D < 0$
27	115.7	223.8	108. 1	16	4.5	3.9	-0.6
25	16.4	102.3	85.9	55	11.3	8.8	-2.5
65	19.3	70.7	51.4	39	59.3	54.1	-5.2
34	20.8	41.1	20.2	20а	26.8	8.7	-18.1
8	9.1	29.5	20.4	20	36.3	8.3	-28.0
13а	8.3	18.8	10.5	66	78.6	49.0	-29.6
56	25.3	34.5	9.2				
9	5.8	12.6	6.8	Выборочные характеристики			
15	0.2	5.5	5.3	Среднее	28.9	42.9	14.0
45	6.6	9.0	2.4	Медиана	19.3	29.5	5.4
21	47.6	49.3	1.7				

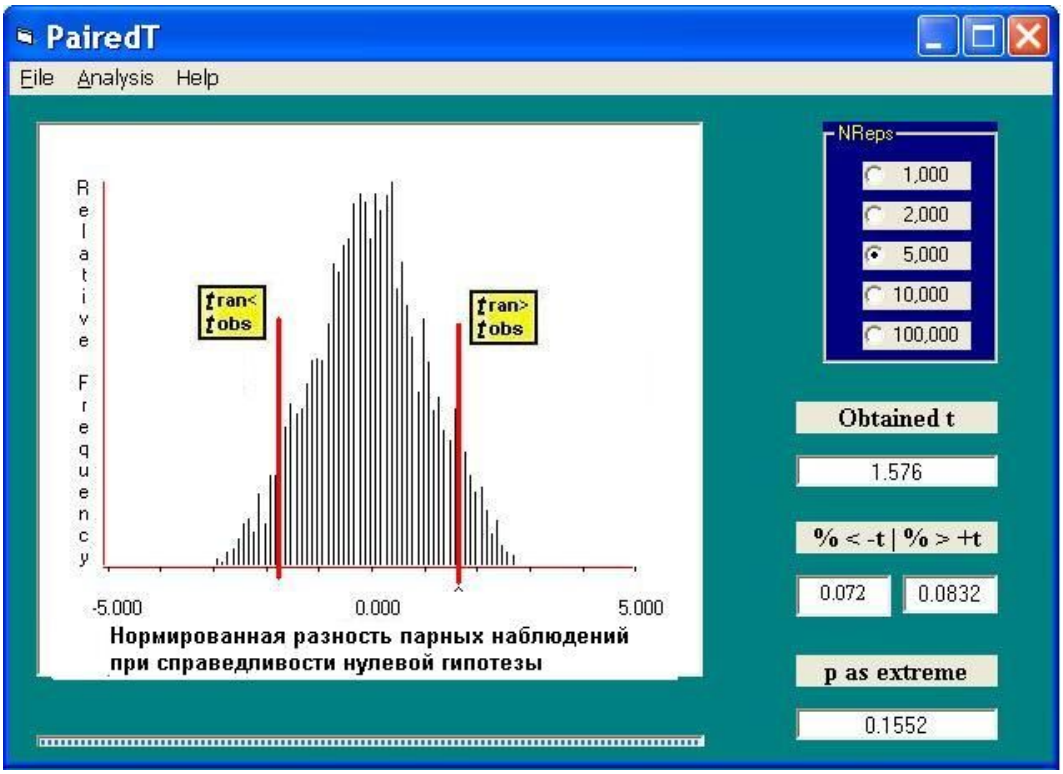


Рис. 2.8. Распределение *t*-статистики, оценивающей изменение численности сине-зеленых водорослей в Куйбышевском водохранилище в 1974-84 гг при справедливости нулевой гипотезы.

Для рандомизации небольших связанных выборок ($n < 25$) Б.Манли (Manly, 2007) предложил алгоритм полного перебора всех возможных знаков разностей, т.е. рассматривается 2^n комбинаций элементов вектора **D**, каждый из которых может принимать значения ‘+’ или ‘-‘. Для каждой сгенерированной комбинации определяется сумма разностей d_i , после чего для оценки *p*-значения подсчитывается число вариантов, у которых эта сумма по абсолютной величине превысила эмпирическое значение.

При анализе данных по сине-зеленым водорослям ($n = 17$) полный перебор включал $2^n = 131072$ итераций, причем в 19386 случаях ($p = 0.148$) имитированное значение суммы разностей находилось дальше от 0, чем для эмпирических выборок. В этих же условиях 9694 нуль-модельных комбинаций ($p = 0.074$) дали положительную сумму разностей d_i, большую, чем при реальном мониторинге, поэтому нет формальных оснований делать вывод о возрастании обилия этой группы фитопланктона.

Если рассматривать полученные результаты в свете методики рандомизации, то можно обратить внимание на то, что по сравнению с опциями программы Resampling 1.3 Ховела (рис. 2.8) мы при использовании полного перебора Манли вместо *t*-статистики использовали простую сумму разностей и одновременно в 26 раз увеличили количество итераций. Однако при этом получили не столь серьезно различающиеся *p*-значения. Таким образом, можно полагать, что *t*-статистика, средняя разность и сумма разностей при рандомизации являются эквивалентными критериями. Не лишним также отметить, что все методы ресамплинга, не использующие схему полного перебора, дают в одних и тех же условиях несколько различающиеся итоговые результаты, причем степень этих различий зависит от наличия экстремальных значений в структуре данных.

С анализом сопряженных значений тесно связан одновыборочный тест Фишера. Пусть мы имеем серию наблюдений x_1, x_2, …, x_n из некоторого неизвестного распределения и предполагается, что заданная величина θ является параметром положения этой выборки. В общепринятых обозначениях проверяется гипотеза H_0: μ = θ, где θ – гипотетическое среднее, которое может быть оценено по другим выборкам или назначено из практических соображений. В определенном смысле, это одновременно и тест на симметричность выборки относительно θ, т.е распределение вероятностей положительных и отрицательных разностей равно: $F(\theta + x) = 1 + F(\theta - x)$.

Проверить нулевую гипотезу H_0: μ = θ можно с использованием теоретического распределения Стьюдента: если не отклоняется гипотеза о нормальном законе распределения наблюдений, вычислить статистику $t = \dfrac{(\bar{x} - \theta)\sqrt{n}}{\sqrt{\sum_n (x_i - \bar{x})^2 / (n-1)}}$ и найти соответствующую ей вероятность при (n - 1) степенях свободы. Другой подход заключается в том, чтобы вычислить разности между наблюдаемыми значениями и гипотетической средней θ, после чего сравнить сумму этих отклонений с распределением, полученным рандомизацией знаков разностей.

В качестве примера используем вариационный ряд из 24 значений концентрации железа (мг/л) в придонном слое воды на 4 станциях Куйбышевского водохранилища в течение 1978 г.: {0.05, 0.06, 0.09, 0.1, 0.11, … , 0.4, 0.43, 0.47, 0.52, 0.6}. По нормам СанПиН 2.1.4.1074-01 содержание железа общего допускается не более 0,3 мг/л, поэтому поставим задачу оценить статистическую значимость отличий результатов мониторинга от этого норматива. Рассчитанные 95% доверительные интервалы концентрации железа составляют $0.185 \div 0.314$ при среднем $\bar{x} = 0.249$, а значение критерия Стьюдента для разностей $t = -1.631$. Вероятность ошибки вывода о превышении нормы с использованием параметрического теста составила $p = 0.058$ при формулировке односторонней гипотезы H_0: μ = θ против альтернативы H_0: μ > θ. При использовании скрипта R по схеме полного перебора Манли после 16777216 итераций выяснилось, что 992628 рандомизированных сумм разностей превысило аналогичную сумму (x_i - θ) для наблюдаемой выборки, т.е. *p*-значение для проверки сформулированной нами гипотезы также оказалось равным 0.059.

К разделу 2.4:

```
# Рандомизационный тест для связанных выборок
source("print_rezult.r")  # Загрузка функций вывода результатов
# Определяем данные по Куйбышевскому водохранилищу (табл. 2.2)
bgw <-
matrix(c(115.7,16.4,19.3,20.8,9.1,8.3,25.3,5.8,0.2,6.6,47.6,4.5,11.3,59.3,26.8,36.3,78.6,
223.8,102.3,70.7,41.1,29.5,18.8,34.5,12.6,5.5,9,49.3,3.9,8.8,54.1,8.7,8.3,49), 17, 2)
colnames(bgw) = c('1974-79', '1980-84')
n <- nrow(bgw)
#  Используем ранговый тест Вилкоксона-Манна-Уитни
wilcox.test(bgw[,1], bgw[,2], paired = TRUE)
#  Или стандартную функцию расчета статистики Стьюдента t.test(...)
```

```
# Возможны значения tail из списка c("two.sided", "less", "greater")
(res = t.test(bgw[,1], bgw[,2], paired=TRUE, alternative="two.sided"))
t.ref =  res$statistic  # t-статистика для эмпирических данных
#--------  Метод 1  --------------
# Рандомизационный тест с заданным числом пермутаций
nperm <- 5000 ; vec.by.rows = as.vector(t(cbind(bgw[,1], bgw[,2]))) ; boots <- numeric(nperm)
for(i in 1:nperm) {
       vec.perm = rep(NA,2*n)
       for(j in 1:n) {
              i1 <- 2*(j-1)+1 ; i2 <- 2*j ; vec.perm[i1:i2] <- sample(vec.by.rows[i1:i2],2) }
       mat = matrix(vec.perm, n, 2, byrow=TRUE)
       res.perm = t.test(mat[,1], mat[,2], paired=TRUE, alternative="two.sided")
       boots[i]  = res.perm$statistic }
RandRes (t.ref, boots, nperm)
#------------  Метод 2  ----------
# Рандомизационный тест с полным перебором знаков разностей (Manly, 2007)
Permdiv <- function (d) {
#  Параметр d  -  разности между парами значений в связанной выборке
#  или разности между выборкой и параметром при одновыборочном тесте
n <- length(d) ; nperm = 2^n
if (n > 30)  {stop ("Перебор займет слишком много времени")}
sum.r <- sum.d <- sum(d) ; asum.d <- abs(sum.d); p <- as.numeric(rep(0,3))
for(j in 1:nperm)  {
       if (sum.r <= sum.d) p[1] <- p[1]+1
       if (sum.r >= sum.d) p[2] <- p[2]+1
       if (abs(sum.r) >= asum.d) p[3] <- p[3]+1
  for (i in 1:n)  if (j %% 2^(i-1) == 0) d[i] <- 0-d[i]
             sum.r <- sum(d)  }
       p <- p/nperm ;  c(cat("Всего итераций B = ",nperm,"\n"),   #  Вывод результатов
  cat("Доля T(rand) < T(obs) P1 = ",p[1],"\n"), cat("Доля T(rand) > T(obs) P2 = ",p[2],"\n"),
  cat("Доля ABS(T(rand)) > ABS(T(obs)) P3 = ",p[3] ,"\n"))
}
Permdiv (bgw[,1]-bgw[,2])                    # Выполнение расчетов
```

2.5. Проблема множественных сравнений

Во многих исследованиях возникает необходимость сравнить две или более выборки не по одному, а по целому комплексу зарегистрированных показателей. Здесь возможны два подхода, связанных с применением: (а) многомерных методов, которые могут оценить различия между группами по всему множеству переменных с учетом их информативно значимых комбинаций, или (б) совокупности тестов, выполняемых для каждой переменной в отдельности, но включающих определенные механизмы, корректирующие уровни значимости статистических гипотез с учетом множественности сравнений.

При проведении статистического анализа данных по большому количеству гипотез необходимо не только выяснить достигнутый уровень значимости в каждом отдельном случае, но и оценить некую глобальную меру ошибки, учитывающую число гипотез. Предположим, мы провели тестирование $m = 6$ индивидуальных показателей для двух групп и в каждом случае достигли уровня значимости различий $p = 0.05$. Однако групповая вероятность (familywise error rate) ошибиться *хотя бы при одном сравнении* значительно превышает 5% и составляет для независимых испытаний $p_F = 1 - (1 - 0.05)^m$, т.е. будет равна 26.5%. В этой ситуации необходимо использовать меньший критический уровень значимости, то есть различия между группами можно считать статистически значимыми, если для каждого из шести показателей справедливо условие $p_F < 0.00851$.

Контроль над групповой вероятностью ошибки p_F на уровне α означает, что мы должны выбрать такие уровни значимости $\alpha_1, ..., \alpha_m$, на которых необходимо проверять гипотезы $H_1, ..., H_m$, чтобы обеспечивалось условие $p_F \leq \alpha$. В частности, метод Бонферрони

заключается в присваивании $\alpha_1 = \ldots = \alpha_m = \alpha/m$. Иными словами, если гипотезы H_i, $i = 1, \ldots, m$, отвергаются на достигнутых уровнях значимости $p_i \le \alpha/m$, то $p_F \le \alpha$.. Аналогичный смысл имеет формула Данна-Шидака (Dunn-Sidak test) $\alpha = 1 - 0.95^{1/m}$.

Однако при увеличении m свыше 6 в результате применения поправки Бонферрони мощность статистической процедуры резко уменьшается, т.к. шансы отклонить неверные гипотезы резко падают. Менее консервативные результаты дает нисходящая процедура Холма, для реализации которой нужно построить вариационный ряд достигаемых уровней значимости $p_1 \le p_2 \le \ldots \le p_m$ для соответствующих нулевых гипотез. После этого выполняется следующая последовательность шагов:

1. Если $p_1 \ge \alpha_1 = \alpha/m$, принять все нулевые гипотезы $H_1, \ldots, H_m$ и остановиться; иначе отвергнуть H_1 и продолжать.
2. Если $p_2 \ge \alpha_2 = \alpha/(m-1)$, принять нулевые гипотезы $H_2, \ldots, H_m$ и остановиться; иначе отвергнуть H_2 и продолжать.
3. Выполнить те же действия для $\alpha_3 = \alpha/(m-2)$, $\alpha_4 = \alpha/(m-3)$, ... , $\alpha_m = \alpha$.

Другая проблема применения поправки Холма-Бонферрони – требование независимости тестируемых выборок: если проверяемые переменные сильно коррелируют между собой, то либо все тесты приведут к значимому результату, либо ни для одного из показателей нулевая гипотеза не будет отклонена. Для этих случаев следует применять байесовские методы.

Рассмотрим различия в размерных характеристиках новорожденных детенышей ящерицы живородящей *Zootoca vivipara*, полученных от самок обычного окраса (выборка 1, $n_1 = 71$) и самок-меланистов (выборка 2, $n_2 = 21$), окрашенных в темно-серые и темно-коричневые тона [пример П5]. Учитывались 4 показателя: M – масса тела (г), L – длина туловища от кончика морды до клоакального отверстия, L_{cd} – длина хвоста, $L_{общ}$ – общая длина с хвостом. Результаты тестирования представлены в табл. 2.3.

Таблица 2.3. Сравнение выборочных средних по четырем размерным показателям детенышей ящерицы, полученных от обычных самок и самок-меланистов.

Показатель	Выборка	Средние m	Рандомизация			По t-критерию	
			$d_{obs} = (m_1 - m_2)$	$P(\lvert d_{ran}\rvert > \lvert d_{obs}\rvert)$	95% доверительный интервал	p для t-критерия	95% доверительный интервал
Масса тела (г)	1	0.218	-0.0186	0.0004	-0.027 ÷ -0.01	10^{-7}	-0.027 ÷ -0.0098
	2	0.236					
Длина тела (мм)	1	20.61	-0.7136	0.001	-1.115 ÷ -0.344	10^{-5}	-1.1 ÷ -0.326
	2	21.33					
Длина хвоста	1	23.94	-1.961	0.0002	-2.75 ÷ -1.14	10^{-5}	-2.796 ÷ -1.126
	2	25.9					
Общая длина	1	44.56	-2.674	0.0002	-3.63 ÷ -1.69	10^{-7}	-3.68 ÷ -1.66
	2	47.23					

Очевидно, что в нашем примере множественная нулевая гипотеза об отсутствии различий между выборками по всему комплексу размерных показателей отвергается при использовании всех методов множественных сравнений. Достигнутые в тесте уровни значимости существенно меньше критических $\alpha = 0.05/4 = 0.0125$ для метода Бонферрони или $\alpha = 0.0127$ для формулы Данна-Сидака, а также не превышают любые значения в последовательности Холма $\{\alpha_1 = 0.0125; \alpha_2 = 0.01667; \alpha_3 = 0.025; \alpha_4 = 0.05\}$.

Доверительные интервалы для разности выборочных средних в случае рандомизации рассчитывались по алгоритму, предложенному Манли (Manly, 2007). Для этого 5000 раз случайно выбранные разности между элементами выборок 1 и 2 прибавлялись к значениям выборки 1, т.е. делалась попытка компенсации различий между выборками. На каждой такой итерации вычислялась сумма имитируемых разностей Δ^*, а также находились значения p_1 и p_2, т.е. доли рандомизированных разностей

выборочных средних, соответственно, меньших или больших этих разностей для эмпирических данных. Интерполяция значений p_1 и p_2 дает нам приблизительные доверительные границы для средней разности между совокупностями. В частности, 95%-ый доверительный интервал задается низким значением Δ^*, которому соответствует p_2 = 2.5 %, и высоким значением Δ^*, приводящим к p_1 = 2.5 %.

К этому можно добавить, что все доверительные интервалы в табл. 2.3 не включают 0, что является также поводом к отклонению нулевой гипотезы. Можно также отметить, что значения доверительных границ, найденных рандомизацией и с использованием *t*-критерия весьма близки между собой.

К разделу 2.5:

```
#   Проблема множественных сравнений
# Определим функцию, корректирующую вектор вероятностей p для проверки
# множественной гипотезы на основе методов Бонферрони-Сидака
#                  Pierre Legendre, May 2007
Sidak <- function(vecP) {
k = length(vecP); vecPB = 0 ; vecPS = 0
for(i in 1:k) {
   bonf = vecP[i]*k ;  if(bonf > 1) bonf=1
   vecPB = c(vecPB, bonf) ;  vecPS = c(vecPS, (1-(1-vecP[i])^k))     }
return(list(OriginalP=vecP, BonfP=vecPB[-1], SidakP=vecPS[-1]))
}
p <- as.vector(c(0.0004, 0.001, 0.0002, 0.0002))
Sidak(p)
```

2.6. Сравнение трех или более независимых выборок

Проверка нулевой гипотезы о равенстве групповых средних H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$ для $k > 2$ выборок может быть выполнена с использованием различных подходов:

- применение однофакторного дисперсионного анализа (ANOVA), в результате которого оценивается совокупная статистическая значимость фактора группировки на основании анализа соотношения дисперсий;
- выполнение всех возможных парных сравнений между группами с использованием апостериорных критериев (post hoc) и проверка сложной (множественной) гипотезы;
- построение линейной модели с группировочным фактором в качестве независимой переменной (см. главу 4).

В качестве примера [П5] рассмотрим три выборки измерений длины тела взрослых самцов живородящей ящерицы *Zootoca vivipara*, отловленных в окрестностях пос. Чепец (n_1 = 35), биостанции Кважва (n_2 = 26) Пермского края и в Кондольском р-не Пензенской обл. (n_3 = 25). Напомним, что дисперсионный анализ основан на трех следующих основных предположениях: (1) независимость выборок (в нашем случае это предположение достаточно очевидно), (2) равенство дисперсии в группах и (3) нормальное распределение изучаемого признака в популяциях, из которых отобраны выборки. Однородность статистической вариации данных в группах будем оценивать на основе статистики Левене (Levene): $L = \dfrac{(n-k)\sum_{i=1}^{k} n_i(\bar{z}_{i\bullet} - \bar{z}_{\bullet\bullet})}{(k-1)\sum_{i=1}^{k}\sum_{j=1}^{n_i}(z_{ij} - \bar{z}_{i\bullet})}$, где $z_{ij} = |x_{ij} - m_i|$, m_i – среднее

или медиана для *i*-й группы, $\bullet$ – символ усреднения по индексу. Тест Левене с использованием медианы показал, что нулевая гипотеза о равенстве дисперсий в группах должна быть отклонена с вероятностью ошибки p = 0.0071.

Стандартная таблица дисперсионного анализа имеет следующий вид (в скобках даны значения *F*-критерия в модификации Уэлча для неравных дисперсий при *df* = 55.2):

Компоненты дисперсии	Сумма квадратов *SS*	Число степеней свободы *df*	Средние квадраты *MS*	*F*-критерий	*p*-значение
Между группами	333.4	2	165.7	7.13 (7.361)	0.0014 (0.00146)
Внутри групп	1930	83	23.25		
Общая	2261	85			

Здесь межгрупповая изменчивость, равная сумма квадратов разностей между средним для каждой группы и глобальным средним $SS_F = \sum_{i=1}^{k} (\bar{x}_{i\bullet} - \bar{x}_{\bullet\bullet})^2$, определяет меру влияния группирующего фактора, а внутригрупповая изменчивость, определяемая суммой квадратов отклонений $SS_e = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_{i\bullet})^2$, – влияние случайных флуктуаций.

Тест на нормальный закон распределения остатков $\varepsilon_{ij} = x_{ij} - \bar{x}_{i\bullet}$ модели дисперсионного анализа $x_{ij} = \bar{x}_{i\bullet} + \varepsilon_{ij}$, выполненный с использованием критерия Шапиро-Уилка, дал положительные результаты (W = 0.98, p = 0.198, гипотеза о нормальности принимается).

Обсудим теперь проблему, как распространить тест рандомизации для двух независимых групп на более общий случай однофакторного дисперсионного анализа при нескольких (k) группах. Очевидно, что нуль-моделью этой структуры данных является случайный порядок размещения элементов по группам, причем принадлежность любого члена к каждой выборке равновероятно. Разумеется мы не будем применять в качестве тестовой статистики сумму значений для первой группы, разность средних для какой-то пары групп или *t*–значение. Необходимо выбрать некий критерий, чувствительный к оценке различий между всеми групповыми средними в совокупности (а значит и к оценке значимости изучаемого фактора). Это может быть, например, сумма квадратов отклонений центров групп от общего среднего SS_F или традиционная F-статистика $F = (n-k)SS_F / (k-1)SS_e$. Очевидно, что при рандомизации они являются эквивалентными тестовыми критериями; но к таковым можно отнести также средние квадраты MS_F или долю $SS_F/(SS_F + SS_e)$, поскольку при любых перестановках изменение всех этих величин происходит совершенно синхронно. В остальных деталях рандомизационная процедура уже вполне знакома по предыдущим разделам:

1. Вычисляем F-значение для эмпирических данных (обозначим как F_{obs}).
2. Генерируем B псевдовыборок, каждая из которых – результат случайной перестановки исходных данных между группами.
3. На каждой иттерации:
   - случайным образом перемешиваем весь комплект данных;
   - назначаем первые n_1 наблюдений в первую группу, следующие n_2 наблюдений во вторую группу, и так далее;
   - вычисляем F_{ran} для этих данных, и, если $F_{ran} > F_{obs}$, увеличиваем на 1 счетчик b.
4. После завершения рандомизации вычисляем $p = (b + 1) / (B + 1)$, которая представляет собой вероятность получения критерия такой же (или более) величины, что и F_{obs}, которая была получена на экспериментальных данных, если верна нулевая гипотеза.
5. Отклоняем или принимаем нулевую гипотезу.

Воспользуемся для анализа представленного примера программой RTAnova (пакет RT – Randomization Testing), разработанной Б. Манли, а также скриптом R, приведенном в дополнении к этому разделу. В результате 5000 итераций рандомизации доля превышения тестовых статистик, полученных на основе случайно полученного разбиения данных по группам, над аналогичным значением для реальных выборок составила p = 0.0026 с

использованием *F*-отношения. Параллельно можно найти и $p = 0.0038$, соответствующее *L*-статистике Левене, хотя подчеркнем, что здесь нам уже не было никакой необходимости акцентировать внимание на нормальности распределения или равенстве дисперсий в группах.

Б.Манли, наряду с обычной схемой рандомизации на основе исходных выборок, использует также алгоритм перестановок с использованием остатков ε_{ij} факторной модели. При его использовании получены результаты, чуть более согласующиеся с выводами, полученными параметрическим методом: $p = 0.0014$ и $p = 0.0024$ для статистик Фишера и Левене соответственно.

Выше с помощью медианного теста была установлена неоднородность дисперсии в группах, которая также хорошо видна на стандартном графике "ящик с усами" рис. 2.9. Поскольку гетероскедастичность обычно может сильно влиять на выводы стандартного ANOVA, представляется важным подтвердить результаты тестирования построением обобщенной линейной модели (GLM), в которой неоднородность дисперсий учитывается соответствующими математическими приемами. Дисперсионный анализ с использованием GLM полностью подтвердил все выводы, сделанные выше параметрическим и рандомизационным тестами. Были рассчитаны информационные критерии Акаике: AIC_F для модели с фактором группировки в качестве независимой переменной и AIC_0 для модели, состоящей из одного свободного члена. Их сравнение с использованием критерия χ^2 привело к $p = 0.0008$ для нулевой гипотезы H_0: $AIC_F = AIC_0$.

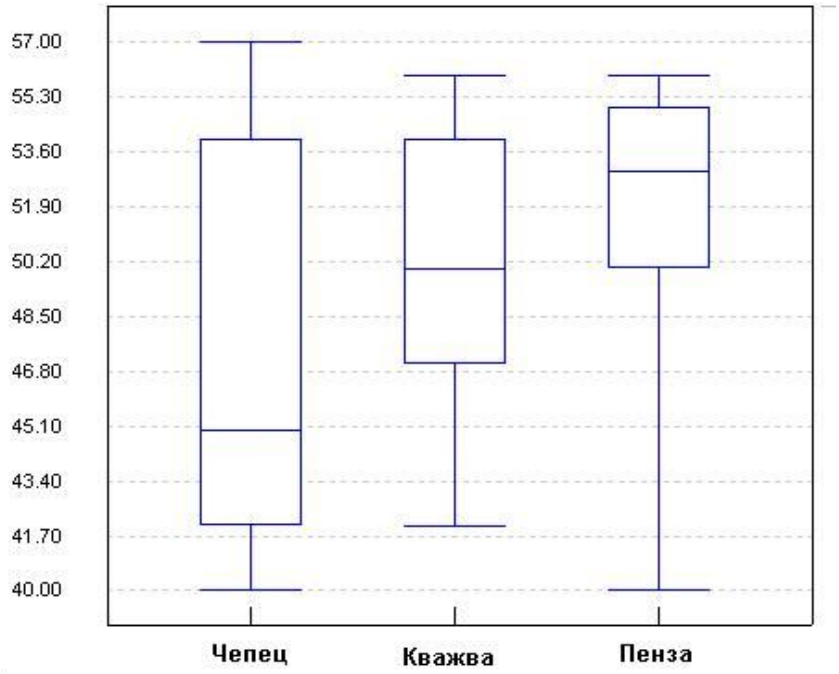


Рис. 2.9. Различия выборочных характеристик длины тела ящерицы живородящей для различных географических регионов

Если по результатам однофакторного дисперсионного анализа отклоняется общая нулевая гипотеза об отсутствии межгрупповых различий, то можно заключить, что не все групповые средние равны. Тогда можно задаться целью установить, между какими конкретно группами ящериц различия в длине тела являются статистически значимыми. Это можно сделать методами множественных сравнений (Post Hoc). Простейший способ – рассчитать для каждой пары выборок *i* и *j* значения обобщенного критерия Тьюки $Q_{ij} = | m_i - m_j | / \sqrt{MS_e / l}$, где *m* – групповые средние, MS_e – средние квадраты отклонений внутри групп, *l* – число сравнений. Из трех возможных частных нулевых гипотез, для которых рассчитаны комбинации *p*-значений {$p_{12} = 0.109$, $p_{23} = 0.245$, $p_{13} = 0.0014$}, соответствующие *Q*-статистикам, отклонена может быть только одна. Иными словами,

статистически значимая географическая вариация данных вызвана лишь отличиями между собой чепецкой и пензенской популяций ящериц, в чем также легко убедиться, проанализировав график на рис. 2.9.

В более общем случае множественные сравнения опираются на теоретические выкладки различных ученых-статистиков и оформляются в виде набора тестов post hoc анализа (например, в пакете SPSS их насчитывается не менее 18). Эти процедуры позволяют исследователю построить итоговые доверительные интервалы, которые можно использовать для попарных сравнений всех средних для всех комбинаций условий. В список наиболее известных тестов включают: проверку наименьшего значения значимой разности LSD, критерий множественного размаха Дункана (Dunkan), метод Стьюдента-Ньюмана-Келса (Student-Newman-Keuls), проверку действительной значимой разности HSD Тьюки (Tukey), а также его альтернативный метод, модифицированный метод LSD, критерий Шеффе (Scheffe), основанный на контрастах, и др. Здесь они перечислены в порядке снижения их мощности (или увеличения консервативности), хотя на этот счет в литературе существуют довольно противоречивые рекомендации.

Для примера [П5] отличие в использовании различных тестов касались лишь гипотезы о равенстве между собой средней длины тела в группах 1 (Чепец) и 2 (Кважа):

Метод	p_{12}	p_{23}	p_{13}
° Бонферрони	0.114	0.382	0.001
° Шеффе	0.115	0.310	0.0016
° Тьюки	0.094	0.277	0.001
° LSD	0.038	0.127	0.00035

Другим примером [П6] однофакторного дисперсионного анализа является оценка влияния минерализации воды на структуру липидов в ульве – многоклеточной макроводоросли *Ulva intestinalis* (L.) Link (Chlorophyta). Сформируем три выборки из массовых долей (%) моногалактозилдиацилглицерола (МГДГ) в составе гликолипидов по результатам анализа липидного состава в биологических пробах из малых рек Приэльтонья с различной степенью минерализации: менее 10 г/л (15 проб), от 10 до 20 г/л (25 проб) и свыше 20 г/л (15 проб).

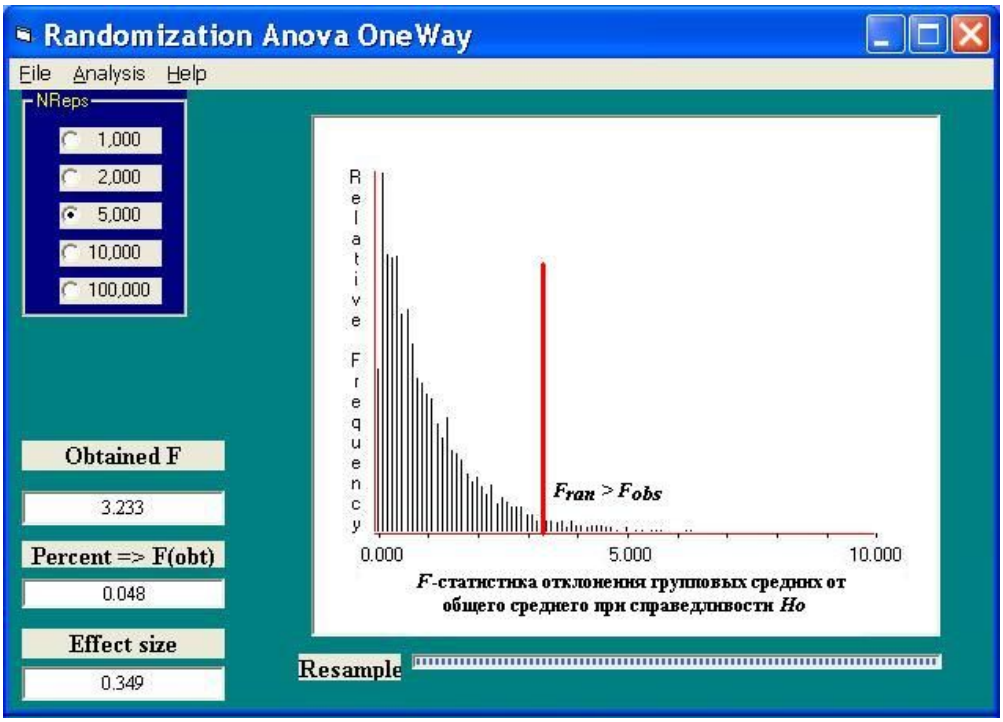


Рис. 2.10. Распределение *F*-статистики, полученное методом рандомизации, для оценки влияния фактора минерализации воды на содержание липида МГДГ в ульве

На рис. 2.10 можно увидеть распределение вероятности значений F, где отмечено местоположение величины F_{obs} = 3.233 для эмпирических данных. Уровень значимости нулевой гипотезы p = 0.048 мы нашли путем подсчета числа итераций ресамплинга с F, большим, чем 3.233. Воспользовавшись стандартной процедурой дисперсионного анализа, можно легко убедиться, что найденное нами p-значение хорошо согласуется с вероятностью, полученной из теоретического распределения Фишера с 2 и 53 степенями свободы (p = 0.0473).

Аналогичный дисперсионный анализ влияния минерализации на содержание общей суммы липидов в тканях *U. intestinalis* (мг/г сырой массы) в тех же условиях дал существенно более веские аргументы отклонить нулевую гипотезу: из 5000 нуль-модельных итераций не было получено ни одной комбинации, для которой имитируемая статистика превысила бы эмпирическое значение F = 14.39 (т.е. p = 0.0002). Хотя постоянно подчеркивается, что p-значение не является «реальной и адекватной мерой статистической убедительности» (Хромов-Борисов, 2011) и их значения в разных опытах никогда не должны сравниваться, в рассмотренном случае легко предположить, что отмеченные сдвиги доли МГДГ в первую очередь являются "вторичным" следствием изменчивости общей массы липидов под влиянием минерализации.

Однако внимательно рассмотрим результаты post hoc анализа. Если использовать тесты множественных сравнений, то все критерии будут утверждать, что особое положение занимает группа 2 с промежуточной минерализацией ($p_{12} = p_{23} < 0.0001$), тогда как различия между 1 и 3-й группами статистически не значимы ($p_{13} > 0.5$) – см. рис. 2.11.

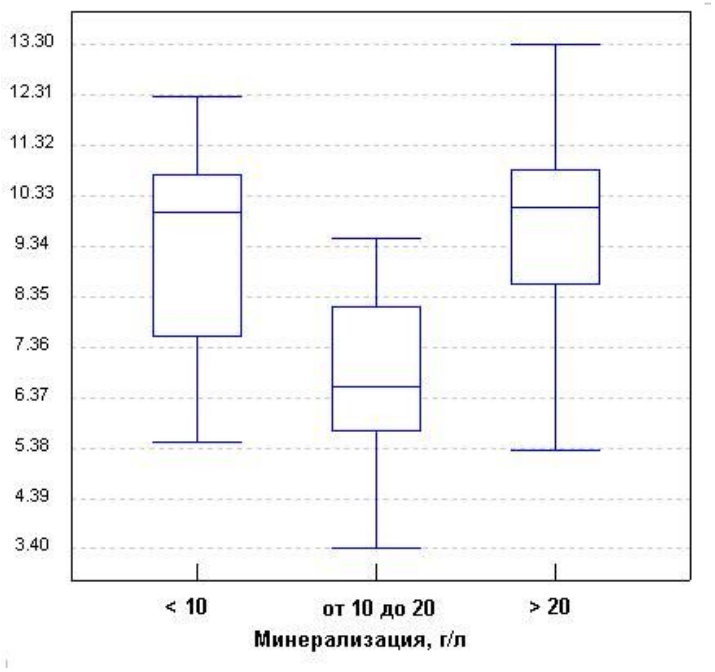


Рис. 2.11. Зависимость общей суммы липидов в тканях водоросли (мг/г сырой массы) от минерализации воды

И здесь можно задаться нетривиальным вопросом: следует ли опираться при статистическом выводе о существовании межгрупповых различий исключительно на тест Фишера или необходимо также привлекать процедуры Бонферрони и его последователей? По крайней мере, при реализации нисходящей процедуры Холма, которую мы подробно обсуждали в разделе 2.3, достаточно одного высокого частного значения p_{13}, чтобы навсегда заблокировать отклонение сложной гипотезы.

Лучший и самый главный совет мы заимствуем из методического пособия для молодых ученых (Советы…, 2005): при применении статистики всегда следуйте мудрому правилу Винни Пуха «Нужно делать то, что нужно, а что не нужно – делать не нужно» ☺.

Но, если всё-таки возникают "недоразумения"[5] при интерпретации результатов статистического анализа, то нужно помнить, что при использовании этих двух подходов играют роль и различная степень консервативности тестов, и разный внутренний смысл проверяемых гипотез.

Другое замечание по существу представленного примера связано с попыткой использовать технику дисперсионного анализа там, где лучше применять иные методы. В частности, ANOVA с комплектацией групп по возрастанию изучаемого количественного показателя (в данном случае – минерализации) существенно уступает по информативности многомерному регрессионному анализу. При применении последнего "странный" феномен снижения содержания липидов в водотоках группы 2 с промежуточной минерализацией может скорее всего оказаться следствием влияния иных факторов, не рассматриваемых при одномерном анализе, и поэтому найти свое корректное биологическое объяснение.

К разделу 2.6:

```
# Однофакторный дисперсионный анализ и перестановочный тест
#  Функции выполняют анализ данных из вектора Y (отклик) и переменной F1, объявленной
as.factor
#  (например, двух произвольных столбцов таблицы)
#  для несбалансированных данных (число повторностей в группах неравно)
anova_ub.1way <- function(Y, F1,  nperm=1000) {
       mat = as.data.frame(cbind(Y,F1)); colnames(mat) <- c("Y","F1"); n = nrow(mat)
       #  Выполнение дисперсионного анализа на эмпирических данных
       F_emp <- oneway.test(Y~F1, data=mat, var.equal=F)$statistic ; GE = 1
       #  Рандомизационный тест F-критерия путем случайной перестановки строк Y
       for(j in 1:nperm)    {
           F_perm = oneway.test(sample(Y,n)~F1, data=mat, var.equal=F)$statistic
           if(F_perm >= F_emp) GE = GE + 1 }
       return (list(oneway.test(Y~F1, data=mat, var.equal=F),
                        paste("P (рандомзация) = ",GE/nperm))) }
#  Выполнение дисперсионного анализа длины тела ящериц  из разных регионов
Len.body <-
c(40,40,40,40,41,41,41,42,42,43,43,43,44,44,44,45,45,45,45,46,47,50,50,51,52,52,54,55,
55,55,56,56,57,57,57,50,46,42,56,50,46,55,54,54,54,47,55,47,52,51,55,50,48,47,55,49,45,
54,46,47,45,51,50,55,55,55,53,50,54,40,54,53,52,55,55,56,56,54,54,53,48,50,50,50,48,51)
Region <- as.factor(c(rep("Чепец",35), rep("Кважва",26), rep("Пенза",25)))
mat = as.data.frame(cbind(Len.body,Region))
# Проверка равенства дисперсий в группах
library(lawstat)
levene.test(Len.body, Region,location="median")
## Проверка нормальности распределения данных в каждой группе по критерию Шапиро-Уилка
shapiro.test.all <- matrix(data=NA, ncol=2, nrow=length(levels(Region)),
dimnames = list(paste("Region", levels(Region), sep = "="),c("W", "P")))
for ( j in 1:length(levels(Region)) ) {
       shapiro.test.each <- shapiro.test(Len.body[Region==levels(Region)[j]])
       shapiro.test.all[j,1] <- shapiro.test.each$statistic
       shapiro.test.all[j,2] <- shapiro.test.each$p.value }
shapiro.test.all
#  Результаты однофакторного дисперсионного анализа с рандомизацией:
(ant <- anova_ub.1way (Len.body,Region))
#  Построение линейной модели с фактором в качестве независимой переменной
m1 <- glm(Len.body ~ Region); summary(m1)
m0 <- glm(Len.body ~ 1) ; summary(m0); anova(m1, m0,test="Chisq")
#  Выполнение множественных сравнений
library(asbio); RF <- as.numeric(Region); RF <- as.factor(RF)
Pairw.test (y=Len.body,x=RF,method="LSD")     ##  LSD метод
Pairw.test(y=Len.body,x=RF,method="Bonf")     ##  Бонферрони
Pairw.test(y=Len.body,x=RF,method="Scheffe")  ##  Шеффе
```

[5] Нам известен оппонент, традиционно указывающий на необходимость применения поправки Бонферрони почти в каждой рецензируемой диссертации

2.7. Преобразование данных

При статистическом анализе часто существуют веские практические соображения ввести в рассмотрение вместо наблюдаемой случайной величины $\boldsymbol{x}$ некоторую функцию $\boldsymbol{y} = g(\boldsymbol{x})$, выполняющую трансформацию выборочных данных. Функциональные преобразования результатов наблюдений выполняются для решения следующих проблем (Кендалл, Стьюарт, 1976, с. 129):

а) сглаживания и "нормализации" распределения исходного вариационного ряда или улучшения выборочных статистических характеристик (универсальные преобразования);

б) приведения к нормальному виду остатков статистической модели, как того требуют параметрические методы (дисперсионный, корреляционный, регрессионный анализ и др.)

в) стабилизации дисперсии в данных, гипотеза о нормальном законе распределения которых не подтверждается; при этом меры положения и рассеяния (масштаба) становятся независимыми от изучаемого параметра θ;

г) устранения взаимодействий факторов, связанных с эффектом нелинейности шкал измерения, после чего можно принять гипотезу об аддитивности эффектов.

Ставя задачу более формально, необходимо попытаться найти такую функцию $g(\boldsymbol{x})$, которая как можно лучше удовлетворяет предположениям статистической модели, оставаясь при этом равным образом тем же множеством "наблюдений", что и $\boldsymbol{x}$. Наиболее распространены нормализующие преобразования для частот (долей): это различные варианты угловых φ-трансформаций $\arcsin\sqrt{p}$ и преобразований Фримана-Тьюки.

Для числовых шкал характерны преобразования, которые относятся к семейству трансформаций по показателю степени (power transformation). Можно, например, выполнить то или иное универсальное преобразование вариационного ряда простыми функциями типа: $\ln(x+1)$, $1/x$, $\sqrt{x}$, $1/\sqrt{x}$, $\sqrt{\ln(x)}$ и т.д. На практике, однако, может оказаться, что использование квадратного корня недостаточно эффективно и не стабилизирует длинный правый "хвост" распределения, тогда как логарифмическое преобразование слишком сильно выражено и приводит к левосторонней асимметрии. Поэтому, если истинное нормализующее преобразование неизвестно, лучшим считается преобразование Бокса-Кокса(Box, Cox, 1964), которое позволяет найти оптимальное решение.

Универсальное семейство преобразований Бокса-Кокса (БК) случайной величины x зависит от значения параметра λ:

- при $\lambda \neq 0$ оно является степенным преобразованием $y(\lambda) = \dfrac{x^{\lambda} - 1}{\lambda}$ с произвольным положительным или отрицательным показателем степени;
- в частных случаях после подстановки параметра λ в основную формулу будем иметь: при $\lambda = -1$ $y = 1/x$; при $\lambda = -0.5$ $y = 1/\sqrt{x}$; при $\lambda = 0.5$ $y = \sqrt{x}$; при $\lambda = 2$ $y = x^2$ и т.д.;
- при $\lambda = 0$ (поскольку деление на нуль приводит к неопределенности) используется логарифмическое преобразование $y(\lambda) = \ln(x)$.

Таким образом, большинство "простых формул" представляют собой лишь частный случай преобразования БК.

Один из способов выбрать наилучшее БК-преобразование – это найти значение λ, доставляющее максимум логарифму функции наибольшего правдоподобия, которая в данном случае имеет вид: $L(x,\lambda) = -\dfrac{n}{2} \cdot \ln\left[\sum_{i=1}^{n} \dfrac{(y_i - \bar{y})^2}{n}\right] + (\lambda - 1)\sum_{i=1}^{n} \ln(x_i)$,

где $\bar{y}$ – среднее для преобразованных данных.

В качестве примера [П2] выполним сравнение численности макрозообентосных организмов (экз/м^2) по результатам гидробиологических проб, сделанных в четырех реках Самарской обл.:

Река	Б. Кинель	Б. Черемшан	Маза	Муранка
Число проб	20	19	9	8
Средняя численность	5281	2119	33435	7499
Ошибка среднего	1044	454	23273	3341
Медиана	3530	1260	10880	4480
Коэффициент асимметрии	1.31	1.59	2.85	1.78

Совокупное для всех 56 проб выборочное распределение численности донных организмов имеет ярко выраженную левостороннюю асимметрию, а использование целого набора критериев согласия различных типов (Крамера-Мизеса, Андерсона-Дарлинга, Шапиро-Уилка, Жарка-Бера, Эппса-Палли, Колмогорова, Смирнова и др.) вызывает единодушное отклонение гипотезы о нормальности исходных данных и остатков дисперсионной модели с высоким уровнем значимости ($p \cong 0$) – см. рис. 2.12. Тест Левене на гомогенность дисперсии в группах также отклонил нулевую гипотезу: p = 0.0002 при использовании средних и p = 0.044 на основе медиан. Можно также отметить аномально высокое значение численности бентоса в одной из проб на р. Маза (217000 экз/м^2, что почти в десять раз больше следующего по величине значения), которое можно квалифицировать как "выброс". Однако наблюдаемые конспецифичные скопления червей трубочника *Tubifex tubifex* – достаточно распространенное в природе явление, чтобы бороться с ним "статистическими методами".

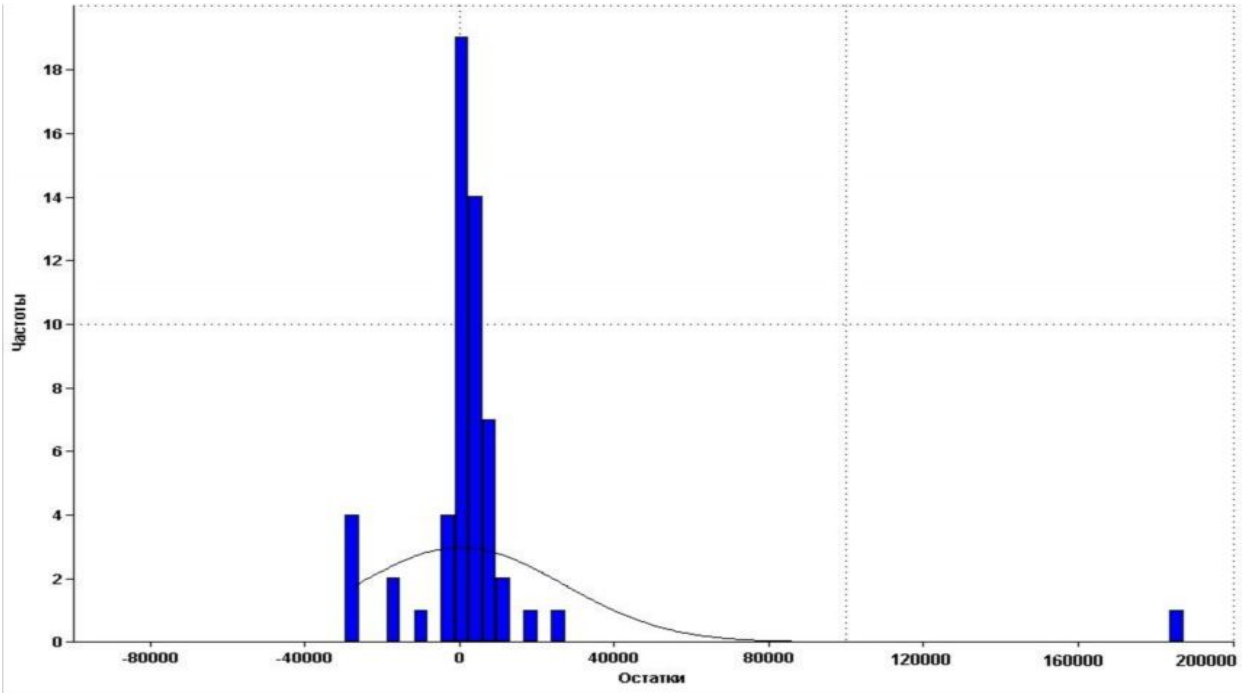


Рис. 2.12. Гистограмма распределения остатков дисперсионной модели для численности донных организмов в четырех реках

Значение показателя степени λ преобразования Бокса-Кокса может быть найдено обычными методами нахождения экстремума на основании двух исходных предпосылок: из условия независимости элементов выборки ($\lambda_{опт}$ = 0.054) и с учетом зависимости численности организмов *Y* от значения фактора River, т.е. водотока, где проводился отбор проб. Для второго случая нахождение оптимального значения $\lambda_{опт}$ = 0.101, соответствующего максимуму функции правдоподобия, представлено на рис. 2.13.

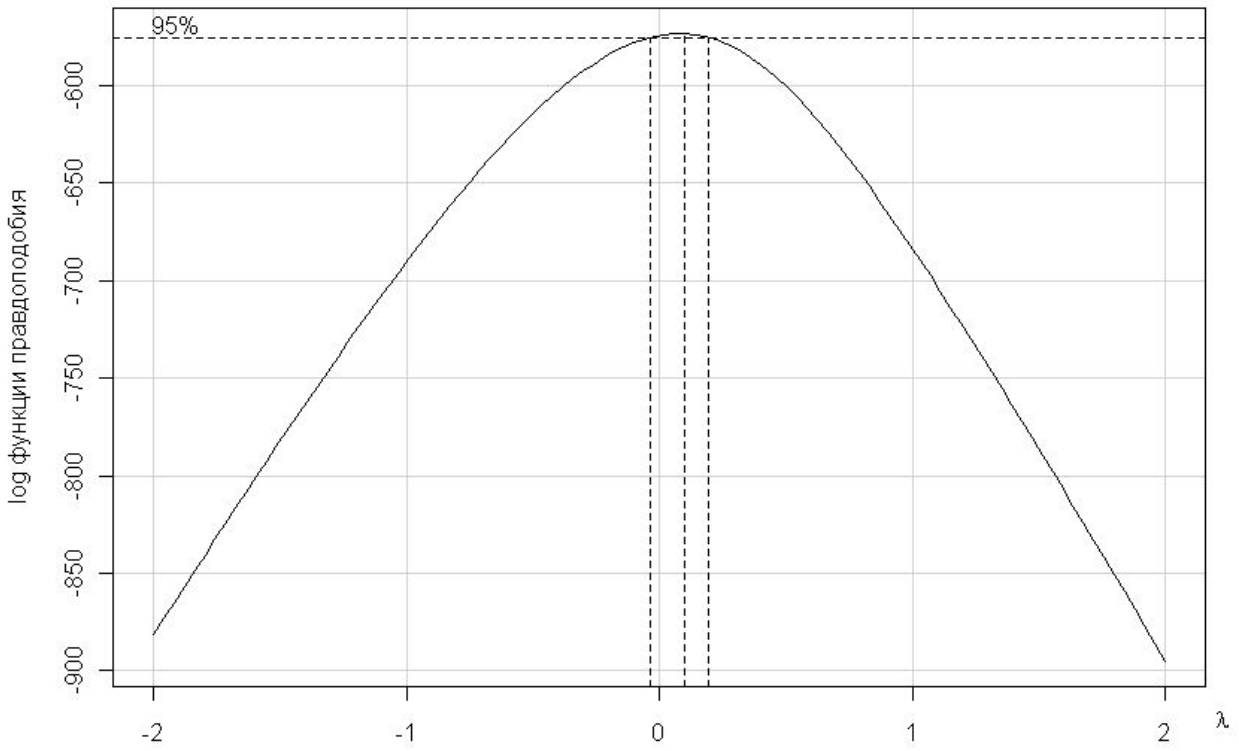


Рис. 2.13. Нахождение оптимального параметра λ преобразования Бокса-Кокса

Тест Левене показал однородность случайной вариации в группах с БК-преобразованной численностью организмов (p = 0.118 на основе медиан). Проверка на нормальность закона распределения с использованием тех же критериев согласия уже не позволила отклонить нулевую гипотезу. В частности, тест Шапиро-Уилка (W = 0.961, p = 0.071) подтвердил нормальность распределения остатков относительно средних в группах после БК-преобразования – см. диаграмму на рис. 2.14.

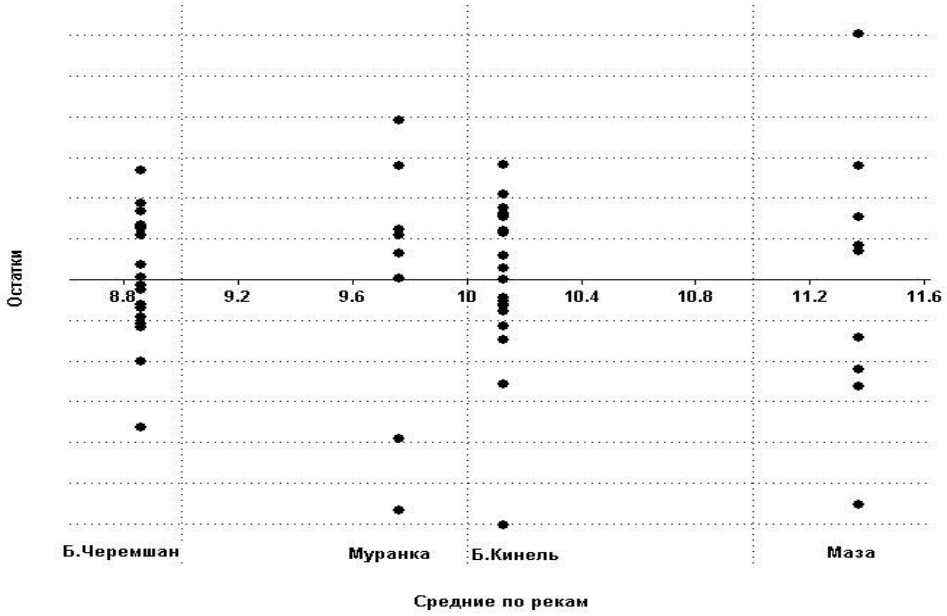


Рис. 2.14. Вариация численности донных организмов относительно групповых средних для четырех рек после преобразования Бокса-Кокса

Несмотря на "революционные" изменения, касающихся исходных предпосылок дисперсионного анализа, непосредственный статистический вывод об отличии популяционной плотности донных организмов в четырех обследованных реках,

сделанный с использованием данных до и после БК-преобразования, остался неизменным. Более того, оценки статистической значимости по *F*-критерию оказались очень близкими, а преобразование данных отразилась только на результатах рандомизационного теста:

Тип данных для расчета	*F*-критерий	Оценка *p*-значения	
		параметрическая	рандомизацией
Натуральные данные, экз/м^2	2.84	0.0468	0.0198
Данные после БК-преобразования	2.80	0.049	0.0469

Если выполить post hoc-сравнения и оценить по обобщенному критерию Тьюки статистическую значимость отличий между каждой парой водотоков, то отклонить H_0 можно только для наиболее полярного случая "Б.Черемшан - Маза" (см. рис. 2.14): $p = 0.036$ на исходных данных и $p = 0.036$ после преобразования Бока-Кокса. Все это свидетельствует об удивительной устойчивости параметрических тестов, проявленной в условиях, когда не соблюдается ни одна из предпосылок дисперсионного анализа.

Было показано (Орлов, 2007), что при больших объемах выборок требование нормальности ослабевает, а при близком объеме выборок не требуется так же и равенства дисперсий. Другими словами, если объемы двух выборок достаточно велики (не менее нескольких десятков) и равны, то проверка равенства математических ожиданий с помощью критериев Стьюдента/Фишера дает правильные результаты не зависимо от того, выполнены ли предпосылки нормальности и равенства дисперсий или нет.

Что касается самой сути идеи трансформации, то необходимо отметить, что, преобразовав данные, мы получаем фактически *другой* показатель, т.е. для практических целей окончательные результаты анализа должны быть выражены только в терминах функции $g(x)$. Переносить эти выводы на исходный показатель мы можем только на уровне "больше - меньше", "растет - падает", поскольку преобразования являются монотонными. Но распространять результаты проверки значимости нулевых гипотез в отношении преобразованной переменной на соответствующие научные предположения о поведении исходной наблюдаемой величины является некорректным. Иными словами, в нашем случае мы можем сделать два возможных заключения, которые не должны пересекаться между собой:

а) имеются статистически значимые различия в реках по средней плотности донного населения в экземплярах/м^2 при несоблюдении исходных предпосылок дисперсионного анализа, или

б) имеются статистически значимые различия в реках по средней преобразованной численности бентоса, выраженной в {(число особей)$^{0.101}$+1}/0.101 на м^2, но при этом предположения дисперсионного анализа соблюдаются.

К разделу 2.7:

```
# Однофакторный дисперсионный анализ и преобразование данных
# Данные о численности макрозообентоса в четырех малых реках
r1 <- c(40,1500,9750,18510,6640,1200,2140,8960,8620,6860,2100,4660,2420,8560,
              560,1920,11860,2260,3840,3220)
r2 <-
c(660,1440,620,860,4180,1780,1260,100,740,3400,4760,2860,3300,3180,1180,920,340,740,7940)
r3 <- c(1720,11920,2880,18000,217000,37080,10880,1280,160)
r4 <- c(5640,5120,160,40,2600,3840,27954,14640)
Data <- data.frame(    Y=c(r1, r2, r3, r4),River = factor(rep(c("Кинель", "Черемшан", "Маза",
                      "Муранка"), times=c(length(r1), length(r2), length(r3), length(r4)))))
#  Выполнение дисперсионного анализа на эмпирических данных
m.full <- lm(Y~River, data=Data) ; anova.res = anova(m.full) ; nperm=5000 ; n = nrow(Data)
#  Рандомизационный тест статистик таблицы ANOVA путем случайной перестановки строк Y
   k = nrow(anova.res) - 1 ;  GE = rep(1,k) ; Pperm = c(rep(0,k), NA)
   for(j in 1:nperm)    { Yperm = sample(Data$Y,n)
       toto = lm(Yperm ~ River, data=Data) ;  anova.per = anova(toto)
      for(i in 1:k) { if(anova.per[i,4] >= anova.res[i,4]) GE[i] = GE[i] + 1 } }
   for(i in 1:k) { Pperm[i]=GE[i]/(nperm+1) } ;    anova.res  = data.frame(anova.res, Pperm)
```

```
   colnames(anova.res) = c("Ст.своб.", "Сумма кв.SS", "Сред. кв. MS",
                           "F крит", "P(парам)", "P(ранд)")
(anova.res)
# Формирование null-модели и ее сравнение с моделью на эмпирических данных
m.null <- lm(Y~1,data=Data) ; print(anova(m.null, m.full))
# Парные сравнения на основе критерия "подлинной значимости" Тьюки
TukeyHSD(aov(Y~River, data=Data))
# Тест Шапиро-Уилка на нормальность распределения остатков
shapiro.test(m.full$residuals)
# Преобразования Бокса-Кокса
library(car)
# Поиск максимума функции правдоподобия и построние графика изменения параметра БК-
трансформации для заданной модели
bc.full <- boxCox(m.full,ylab = "log функции правдоподобия" )
 bc.full.opt <- bc.full$x[which.max(bc.full$y)]
print(paste("Оптимальная лямбда БК-преобразования для факторной модели:",bc.full.opt))
# Заполнение таблицы преобразованными данными и построение факторной модели
Data$Y.bc.full <- bcPower(Data$Y, bc.full.opt)
m.bc.full <- lm(Y.bc.full ~ River,data=Data)
#  Тест Тьюки и Шапиро-Уилка на модели с преобразоваммми данными
TukeyHSD(aov(Y.bc.full~River, data=Data)) ; shapiro.test(m.bc.full$residuals)
#  Преобразования Бокса-Кокса для нулевой модели
bc.null <- boxCox(m.null); bc.null.opt <- bc.null$x[which.max(bc.null$y)]
print(paste("Оптимальная лямбда БК-преобразования для нулевой модели:",bc.null.opt))
Data$Y.bc.null <- bcPower(Data$Y, bc.null.opt) ; m.bc.null <- lm(Y.bc.null ~ 1, data=Data)
# Оценка статистической значимости преобразованной факторной модели
print(anova(m.bc.null, m.bc.full))
```

2.8. Сравнение видового разнообразия систем и ограничения на рандомизацию

В разделе 1.6 нами обсуждались проблемы оценки интервальных значений различных индексов, характеризующих многовидовые композиции (которые экологи трепетно именуют "*сообществами*"). Однако использование доверительных интервалов не дает нам точной оценки уровня значимости нулевой гипотезы. В разделе 2.3 этой главы мы также сравнивали изменчивость точечных значений одного из таких индексов видового разнообразия – энтропии Шеннона – для двух участков средней реки Сок [пример П2]. Но сопоставление выборок безликих точечных показателей – это не вполне то же самое, что сравнение двух видовых структур, каждый компонент которой имеет специфическую функцию статистического распределения, а все вместе они определяют сложный характер взаимодействий между элементами внутри сообщества (или, научно выражаясь, эмерджентность экосистемы).

Простейший способ представить видовую структуру сообществ – это сформировать таблицу показателей популяционной плотности (в нашем случае, средней численности организмов на м^2) каждого вида, встреченного на l сравниваемых участках – см. табл. 1.2 в разделе 1.6. Нулевая гипотеза об однородности разнообразия для l анализируемых многовидовых композиций заключается в предположении, что вычисленные эмпирические индексы I равны: H_0: $I_1 = I_2 = ... = I_l$. Например, при сравнении двух сообществ тестируемым статистическим критерием может быть принята разность двух индексов $\tau_{obs} = |I_1 - I_2|$, вычисленных для наблюдаемых данных. Так для верхнего и нижнего участков р. Сок разность индексов Шеннона равна

$$\tau_{obs} = |H_1 - H_2| = |3.20 - 3.28| = 0.08,$$

но насколько велико это значение, чтобы не отвергать нулевую гипотезу?

Предположим, нам удалось получить неизвестное статистическое распределение этой разности при справедливости нулевой гипотезы. Если эмпирическое значение τ_{obs} окажется в критической области нуль-распределения, то нулевая гипотеза будет отклонена на соответствующем уровне значимости.

Как было показано выше, получить частотное распределение плотности условной вероятности статистики принятого критерия в предположении, что нулевая гипотеза верна, можно с использованием различных алгоритмов рандомизации. Если нам удастся корректно сгенерировать большое число значений разностей τ_{sim} на рандомизированных данных, то мы можем подсчитать относительную частоту, с которой имитируемые величины превышают значение разности на данных, которые мы получили в эксперименте. Это и будет искомая оценка достигнутого уровня значимости.

Сгенерировать нуль-распределение можно с использованием соответствующим образом подобранной нуль-модели, т.е. способа имитации структуры наблюдаемых данных, преднамеренно исключающего возможные механизмы влияния рассматриваемого фактора (в нашем случае – пространственной изменчивости биоразнообразия).

Однако так ли просто сформулировать соответствующую задаче систему ограничений и подобрать для ее реализации адекватную нуль-модель? Можно, например, предположить, что на обоих участках одинаковое гамма-разноообразие, а каждый вид имеет характерное для изучаемого региона специфическое частотное распределение. Или, наоборот, сделать предположение, что в обоих сообществах одинаковое бета-разнообразие и шанс встретить особь каждого вида равновероятен.

Ранее мы использовали исключительно перестановочный тест, который использует значения численностей, полученные в эксперименте, но все элементы таблицы каждый раз хаотически и равновероятно перемешивает по всем ее ячейкам – модель 1 на рис. 2.15. Однако, переставляя местами значения численностей, то мы тем самым резко нарушаем баланс частот как для видов, так и для участков. Такая реализация нуль-модели, у которой есть только одно ограничение – постоянная общая сумма всех особей, обычно называют модель **EE**, т.е. Equiprobable или равновероятной. К сожалению, такая модель часто может оказаться не вполне экологически оправданной.

Исходная матрица

пп	Виды	Сок_верх	Сок_нижн	Итого
1	ChEuk.gr	124	0	124
2	ChTar.sp	542	140	682
3	ChPlc.co	495	18	513
4	ChCld.m.	193	353	546
5	ChChi.n.	0	542	542
6	ChLip.a.	0	524	524
7	ChPro.o.	357	0	357
8	EpBst.r.	13	3	16
9	ChCri.b.	360	49	409
	Итого	2084	1629	3713

1. Перестановочная модель EE

Сок_верх	Сок_нижн	Итого
0	0	0
13	353	366
360	542	902
140	542	682
3	0	3
49	357	406
124	0	124
495	193	688
524	18	542
1708	2005	3713

2. Обмен в строках (модель EF1)

Сок_верх	Сок_нижн	Итого
0	124	124
140	542	682
495	18	513
353	193	546
0	542	542
0	524	524
0	357	357
13	3	16
360	49	409
1361	2352	3713

3. Фиксированная по видам (EF2)

Сок_верх	Сок_нижн	Итого
0	124	124
329	353	682
227	286	513
267	279	546
0	542	542
524	0	524
357	0	357
8	8	16
201	208	409
1913	1800	3713

4. Дважды фиксированная модель FF

Сок_верх	Сок_нижн	Итого
68	56	124
371	311	682
285	228	513
298	248	546
299	243	542
307	217	524
194	163	357
9	7	16
253	156	409
2084	1629	3713

Рис. 2.15. Примеры реализации четырех основных нуль-моделей рандомизации

В нашем случае будет более соответствовать смыслу поставленной задачи комбинированная модель 2 **EF**1 (Equiprobable- Fixed), т.е. суммарные численности видов всегда остаются постоянными, но в случайном наборе случайных строк происходит простая перестановка значений между столбцами. Тот же алгоритм обмена заложен и в модели 3 **EF**2, но ненулевые численности тем или иным способом пропорционируются и перераспределяются по ячейкам в строках, принимая случайные целочисленные значения, но с неизменяемыми общими суммами для каждого вида. В обеих моделях суммарная

численность для каждого участка (т.е. по столбцам) является непредсказуемой.

Максимум ограничений на рандомизацию представлен в модели r2dtable (Patefield, 1989) или дважды фиксированной модели 4 **FF** (Fixed-Fixed). Она требует, чтобы общая численность особей каждого вида в строках и суммарная численность для каждого участка в столбцах нуль-матрицы в точности соответствовала бы наблюдаемым значениям в эмпирической матрице. Сами значения численностей в ячейках могут принимать случайные целочисленные значения.

Следует отметить, что придумано еще полтора десятка нуль-моделей комбинированного и пропорционального типов на все случаи экологической жизни (Gotelli, Graves, 1997).

Результаты расчетов могут отображаться в виде графика ядерной оценки плотности распределения тестируемой статистики при справедливости нулевой гипотезы – см. рис. 2.16.

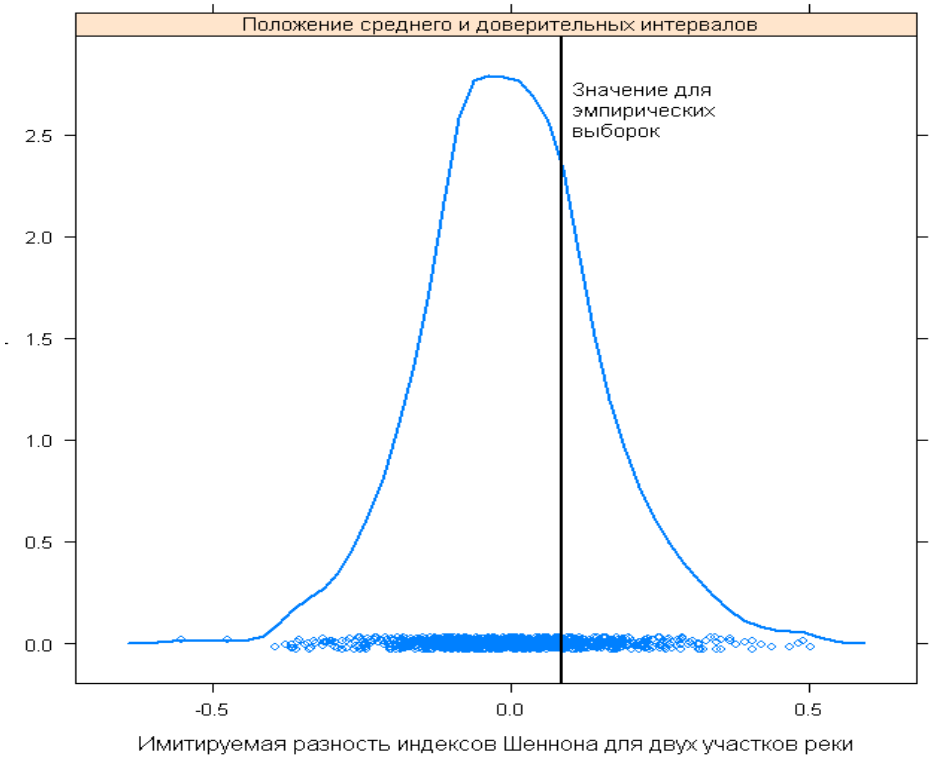


Рис. 2.16. Функция ядерной оценки плотности распределения тестовой статистики по нуль-модели EF1 для сравнения видового разнообразия макрозообентоса двух участков р. Сок

Результаты статистического анализа с использованием описанных выше типов нуль-моделей представлены в табл. 2.4: три использованные модели из четырех оказались синхронны в выводе об однородности видового разнообразия донных сообществ на всем протяжении р. Сок.

Таблица 2.4. Сравнение видового разнообразия двух участков р. Сок с использованием различных типов нуль-моделей в среде R при $\tau_{obs} = |H_1 - H_2| = |3.20 - 3.28| = 0.08$

Тип нуль-модели	95% доверительные интервалы τ_{ran} при справедливости H_0	Вероятность $Pr(\tau_{ran} \geq \tau_{obs})$
1. Перестановочная модель EE	-0.501 ÷ 0.542	0.760
2. Обмен в строках (модель EF1)	-0.290 ÷ 0.313	0.520
3. Фиксированная по видам модель EF2	-0.226 ÷ 0.222	0.485
4. Дважды фиксированная модель FF	-0.066 ÷0.063	0.0089

Можно отметить, что в ряду моделей закономерно сужается доверительный интервал тест-критерия, т.е. каждая следующая модель имеет существенно меньший разброс имитаций индексов Шеннона (и их разностей) и оказывается все менее и менее

либеральной по отношению к нулевой гипотезе. В практическом плане в каждом конкретном случае необходима ориентация на экологически реалистические нуль-модели, которые в состоянии отличать экологические и эволюционные механизмы от чисто статистических и выборочных артефактов (Gotelli, McGill, 2006).

Оценка различий показателей видового разнообразия для трех или более сообществ основана на алгоритмах множественных парных сравнений в рамках моделей однофакторного дисперсионного анализа (ANOVA). Метод бутстреп-оценки одновременных (simultaneous) доверительных интервалов тестовой статистики при сравнении произвольного индекса *I* в нескольких группах описан в работах Р. Шерера с соавторами (Scherer, Schaarschmidt, 2013) и программно реализован в пакете simboot статистической среды R.

Пусть мы имеем матрицу наблюдений из *S* строк (виды) и *K* столбцов наблюдений (пробы или средние численности по местообитаниям). Каждый столбец $j = 1, \ldots, J_i$ относится к *i*-й группе (участку), $i = 1, \ldots, L$. При таком способе группировки по обычным формулам ANOVA можно вычислить:

- значения индекса видового разнообразия I_{ij} для каждой *j*-й пробы;
- групповые средние $\overline{I}_i$ для каждой *i*-й группы проб из *L*;
- остатки $\varepsilon_{ij} = I_{ij} - \overline{I}_i$, групповые средние остатков $\bar{\varepsilon}_i$ и оценку $\hat{\sigma}^2_\varepsilon$ остаточной дисперсии.

Между этими группами возможно *M* множественных сравнений, механизм реализации которых определяется матрицей **C** априорных (т.е. вводимых из общих соображений) ортогональных *контрастов*. Коэффициенты контрастов c_{mi} задают смысл проверяемых гипотез и подчиняются условию $\sum_i^L c_{mi} = 0$ для всех $m = 1, \ldots, M$. Например, матрица контрастов Тьюки $\mathbf{C}_T$ определяет при $L = 3$ все возможные переборы групп, а контрасты Даннета $\mathbf{C}_D$ – механизм сравнения контрольной группы со всеми остальными:

$$\mathbf{C}_T = \begin{pmatrix} -1 & 1 & 1 \\ -1 & 0 & 1 \\ 0 & -1 & 1 \end{pmatrix}; \qquad \mathbf{C}_D = \begin{pmatrix} -1 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix}.$$

Набор *M* из *p*-значений, скорректированных с учетом множественных сравнений между группами (multiplicity-adjusted *p*-values), рассчитывается с использованием бутстрепа следующим образом:

1. Выполняется непараметрический бутстреп вектора остатков ε_{ij}. Случайный выбор с возвращениями осуществляется только в пределах каждой группы из *L*. Вычисляются групповые средние $\bar{\varepsilon}^*_i$ и общая дисперсия остатков $\hat{\sigma}^{2*}_\varepsilon$ на бутстреп-выборке.
2. Для каждого *m*-го варианта парного сравнения групп вычисляется статистика межгрупповых отличий $t^*_m = \dfrac{\sum_i c_{mi}\,\bar{\varepsilon}^*_i}{\hat{\sigma}^{2*}_\varepsilon \sqrt{\sum_i c^2_{mi} / J_i}}$, учитывающая контрасты *i*-й группы.
3. Шаги 1-2 повторяются *B* раз и на каждой *b*-й итерации находится максимум $(t^*_b)_{\max}$ из всех возможных *m*-х вариантов парного сравнения.
4. Пусть $R_{bm} = 0$ при $t_m > (t^*_b)_{\max}$ и $R_{bm} = 1$ в противном случае, где t_m – эмпирическое значение тестовой статистики. Тогда скорректированное *p*-значение для *m*-го контраста с учетом множественных сравнений будет равно $\tilde{p}_m = (\sum_b R_{bm})/B$.

Важно отметить, что, используя описанный алгоритм, мы можем перейти от "кажущегося" разнообразия, оцениваемого по средним численностям видов в результате механического объединения (в нашем случае) 50 выборок, к реальному разнообразию каждой гидробиологической пробы. То, что внутренняя гетерогенность видового состава на каждом из трех участков существует в любых масштабах (створ, поперечный профиль,

точка), не требует доказательств. Это видно, в частности, на мозаичной диаграмме, где вся речная система р. Сок показана с дискретностью 13-ти станций наблюдений – рис. 2.17.

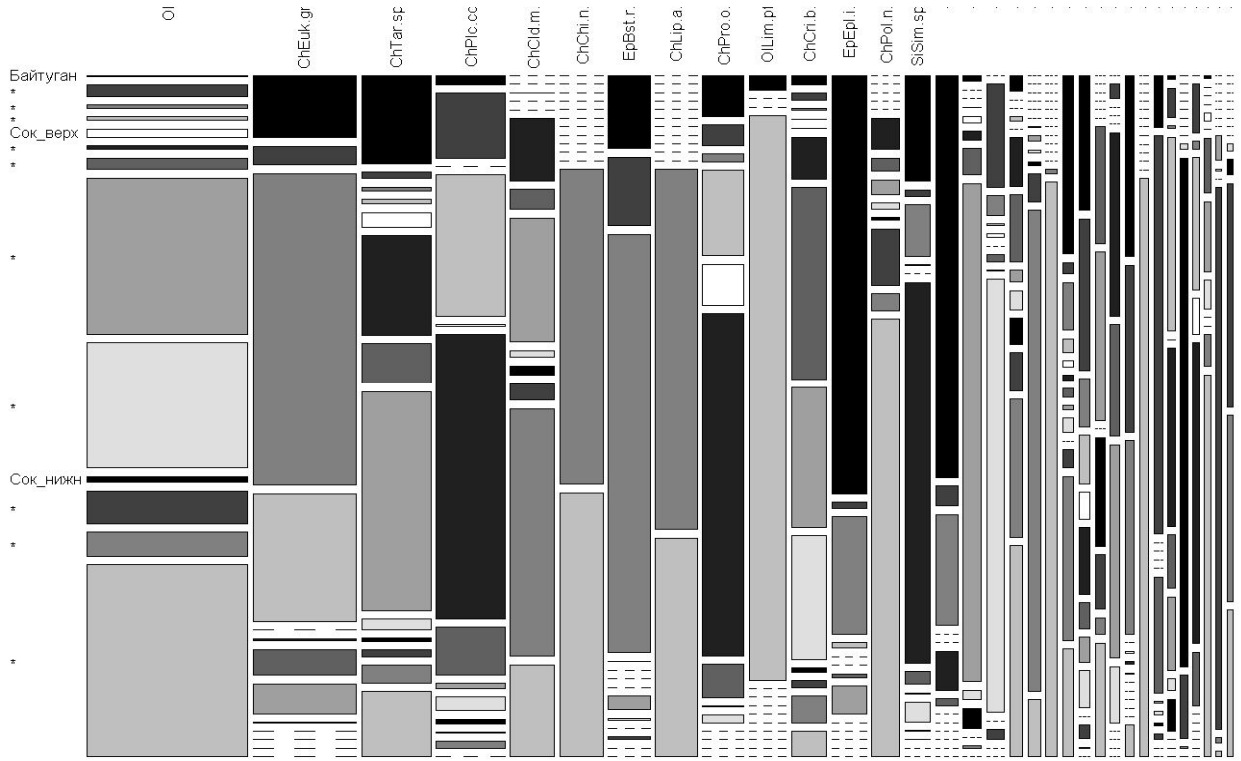


Рис. 2.17. Мозаичная диаграмма относительной численности видов макрозообентоса (представлены вверху кодами базы данных) на 13 станциях наблюдения рр. Сок и Байтуган

В табл. 2.5 представлены результаты сравнения усредненных гидробиологических проб на 13 станциях, сгруппированных по трем участкам. При использовании контрастов Даннета в качестве контрольного участка был взят верхний участок Сок. Как и в вышеприведенных примерах, нулевая гипотеза об однородности уровня видового разнообразия на всем протяжении речной системы не отклоняется.

Таблица 2.5. Сравнение видового разнообразия трех участков р. Сок с использованием функции sbdiv(…) пакета simboot после обработки матрицы 13×374 средних численностей видов на 13 станцияхнаблюдений ; t_m - эмпирическая статистика (аналог *t* Стьюдента); CI_l и CI_u - нижний и верхний доверительные интервалы; *p* и p_{adj} - уровни значимости без учета и с учетом множественных сравнений

Сравниваемые участки	t_m	CI_l	CI_u	*p*	p_{adj}
1. Использование контрастов Тьюки (все возможные переборы)					
Сок_верх - Байтуган	-0.063	-0.886	0.759	0.828	0.976
Сок_нижн - Байтуган	0.150	-0.716	1.017	0.651	0.876
Сок_нижн - Сок_верх	0.214	-0.609	1.036	0.756	0.482
2. Использование контрастов Даннета (контрольная группа – "Сок_верх")					
Байтуган - Сок_верх	0.063	-0.728	0.855	0.838	0.963
Сок_нижн - Сок_верх	0.214	-0.578	1.005	0.477	0.714

К разделу 2.8:

```
#  Нуль-модели и ограничения на рандомизацию
library(vegan)                # Включение пакета vegan
library(lattice)  ## Включение пакета, обеспечивающего вывод графиков
source("print_rezult.r")  # Загрузка функций вывода результатов
#  Загрузка исходных данных для расчета из двоичного файла
load(file="Сок_Байт.RData") ; ls() ; TT
```

```
TT2S <- t(TT[,3:4]) ; TT2S <- TT2S[,colSums(TT2S)!=0] #  Берем только два участка Сок
# Определение функции, вычисляющей три оценки разнообразия:
# видовое богатство, индексы Шеннона и Симпсона
div_est <- function (data, est = "shannon")
       { if (est == "shannon" | est == "simpson" )  diversity(data,est)
        else if (est == "rich")  sum(data>0) }
##  Функция, задающая разность между показателями видового разнообразия
divdiff <- function(x) div_est(x[2, ]) - div_est(x[1, ])
# -------------------              Выполнение расчетов
Nperm <-1000  #   Задаем число рассчитываемых экземпляров модели
# Сравнение показателей видового разнообразия для эмпирических выборок
div_est(TT2S[2, ]) ; div_est(TT2S[1, ]) ; d_emp <- divdiff(TT2S)
# ---------- Анализ нуль-моделей с использованием функции oecosimu()
#      1-й вариант - модель EE
#  Определение функции, выполняющей перемешивание значений в столбцах и строках
rand.all <- function(dat) {
       dat.l<-length(dat[1,]) ; pool <- sample(c(dat[1,],dat[2,]),replace=F)
       dat[1,]= pool[1:dat.l]; dat[2,] = pool[(dat.l+1):(2*dat.l)]; dat }
(null.1 <- oecosimu(TT2S, divdiff, rand.all, nsim = Nperm))
#      2-й вариант - модель FE1
#  Определение функции, выполняющей обмен значений между обеими строками
#  для случайных столбцов в случайном количестве
swap.row<- function(dat) {
       dat.l<-length(dat[1,]) ;  x<-round(runif(1,min=0, max=dat.l))
       if(x>0) tmp<-sapply(sample.int(dat.l, x, replace=FALSE),
              function(i) dat[,i]<<-rev(dat[,i])); dat  }
 (null.2 <- oecosimu(TT2S, divdiff, swap.row, nsim = Nperm))
#  Построение графика ядерной плотности вероятности
densityplot(null.2, xlab="Имитируемая разность индексов Шеннона для двух участков реки",
                                        lwd=2,  ylab="Плотность вероятности")
#      4-й вариант - модель FF (r2dtable)
(null.3 <- oecosimu(TT2S, divdiff, "r2dtable", nsimul = Nperm))
# ------- Анализ нуль-моделей с использованием функции permatfull ()
#   3-й вариант -   модель FE2
d_rand <- rep(0,Nperm)  ; for (i in 1:Nperm) {
       x3 <- permatfull(TT2Sc, fixedmar = "columns", shuffle = "both" , times = 1)
       TT_rand <- as.matrix(x3$perm[[1]]) ; d_rand[i] <- divdiff(TT_rand)  }
RandRes (d_emp, d_rand , Nperm)
# -------  Сравнение трех участков и вычисление скорректированных p-значений
library(simboot)
TTS ; datspec = TTS[,-1] ; datspec = as.data.frame(t(datspec))
Variety <- as.factor(c(rep("Байтуган",4),rep("Сок_верх",5),rep("Сок_нижн",4)))
sbdiv(X = datspec, f = Variety, theta = "Shannon", type = "Tukey", method = "WYht",
              conf.level = 0.95,alternative = "two.sided", R = 2000)
sbdiv(X = datspec, f = Variety, theta = "Shannon", type = "Dunnett", method = "WYht",
              conf.level = 0.95, alternative = "two.sided", R = 2000, base = 2)
# Мозаичная диаграмма
COUNTS<-as.matrix(datspec) ; rownames(COUNTS)<-Variety
COL<-grey(c(0,2,4,6,8,1,3,5,7)/8) # Связь оттенка цвета и градации численностей видов
DMO<-COUNTS[,order(colSums(COUNTS), decreasing=TRUE)] ; DMO <- DMO[,1:33]
colnames(DMO)[15:33]<-"."  ; rownames (DMO)[c(2:4,6:9,11:13)]<-"*"
par(mar=c(4,2,1,1)) ; mosaicplot(t(DMO), col=COL, las=2, off=15, main="", cex=1.1)
```

2.9. Сравнение индексов таксономического и функционального разнообразия

Такие императивные меры оценки разнообразия, как индексы Джини-Симпсона или Шеннона, не столько отражают подлинную видовую структуру сообществ, сколько основываются на «экономности объясняющих принципов» (Левич, 1980). В результате «разнообразие как экологическое понятие стало весьма отличаться от разнообразия как статистического индекса» (Ricotta, 2005). Еще Э. Пиелу (Pielou, 1975) заметила, что нельзя любое многообразие интерпретировать просто как механическую смесь эквивалентных

компонентов, поэтому, например, индекс биоразнообразия для сообщества из таксономически различных видов (орел, сорока и чернозобик) должен быть при прочих равных условиях выше, чем у сообщества из близких между собой популяций (сорока, голубая сорока и древесная сорока). Это вполне естественное соображение вызвало появление ряда новых концепций и постоянно увеличивающееся множество мер разнообразия, учитывающих таксономическое или функциональное дифференцирование видов.

Наиболее целенаправленные попытки учесть в явном виде структурную информацию, необходимую для оценки биологической сложности экосистемы, связаны с квадратичной энтропией Рао (Rao, 1982):

$$Q = \mathrm{PDP}^{\mathrm{T}} = \sum_{i=1}^{S}\sum_{j=1}^{S} d_{ij} p_i p_j \,. \qquad (2.1)$$

Она учитывает как вероятности p_i (или относительные популяционные плотности отдельных видов), так и оценки d_{ij} различий между каждой парой видов, $i, j \in S$, $d_{ii} = 0$, $d_{ij} = d_{ji}$. В зависимости от поставленной задачи исследований компоненты квадратной симметричной матрицы **D** могут интерпретироваться как те или иные экологические расстояния: таксономическая удаленность, генетическая дивергенция (результат сравнения участков последовательностей ДНК, принадлежащие двум генотипам), функциональные отличия и др.

Квадратичная энтропия Рао представляет собой обобщенную форму индекса видового доминирования Симпсона $C = \sum_{i=1}^{S} p_i^2$ и определяет среднее экологическое различие между двумя случайно извлеченными из сообщества особями. При этом индекс Рао может быть разложен на следующие составляющие: $Q = C\,D_{\mathrm{m}} + B$,
где C – индекс Симпсона, D_{m} – среднее экологическое расстояние между видами, B – фактор баланса относительных частот, который рассчитывается как ковариация между d_{ij} и $p_i p_j$, если считать их случайными переменными.

Таксономическое разнообразие принимает во внимание генерализацию богатства видов, основанную на подсчете сумм длин ветвей или числа узлов таксономического дерева, построенного, в первом приближении, на основе линнеевской классификации. Если виды расположить в соответствии с такой иерархией по типам, классам, отрядам, семействам, родам, то меру таксономического различия w_{ij} двух видов i и j можно задать как длину половины пути, который связывает эти виды по ветвям дерева. Например, если два вида принадлежат к одному роду, то нужно пройти один шаг для того, чтобы достичь общего узла, следовательно, $w_{\mathrm{ij}} = 1a$, где a – стандартное расстояние между смежными узлами. Если виды принадлежат к разным родам, но одному семейству, то потребуется два последовательных шага ("вид–род" и "род–семейство") и так далее.

В этих обозначениях индекс среднего таксономического своеобразия (average taxonomic distinctness), предложенный Р.Уорвиком и К.Кларком (Warwick, Clarke, 1995) представляет среднюю длину ветвей w_{ij} между любой парой видов и их общим узлом иерархии:

$$\Delta^{+} = \frac{2}{S(S-1)} \sum_{i=1}^{S-1}\sum_{j=i+1}^{S} w_{ij} \,.$$

Длина шага a была стандартизирована таким образом, чтобы максимальное расстояние w_{ij} было бы равно 100, т.е. значение индекса Δ^{+} изменялось в пределах от 0 до 100.

Если в формуле для Δ^{+} дополнительно учесть относительное обилие видов p_i, то получим индекс таксономического разнообразия (taxonomic diversity) Δ, который является простой модификацией квадратичной энтропии Рао Q, и индекс вариации (дисперсии) таксономического различия Δ^{*}:

$$\Delta = \frac{2}{S(S-1)} \sum_{i=1}^{S-1}\sum_{j=i+1}^{S} w_{ij} p_i p_j \qquad \Delta^{*} = \sum_{i=1}^{S-1}\sum_{j=i+1}^{S} w_{ij} p_i p_j \Big/ \sum_{i=1}^{S-1}\sum_{j=i+1}^{S} p_i p_j \,.$$

Аналогом индекса видового разнообразия Шеннона является таксономическая энтропия (Ricotta, Avena, 2003):

$$H(\mathbf{P},\mathbf{K}) = -\sum_{i=1}^{S} p_i \ln k_i \;;\;\; k_i = \sum_{j=1}^{S} d_{ij} \Big/ \sum_{i=1}^{S}\sum_{j=1}^{S} d_{ij} \;;\; \sum_{i=1}^{S} k_i = 1, \tag{2.2}$$

где $\mathbf{K} = (k_1, k_2, \ldots, k_s)$ – вектор относительных вкладов видов в таксономическое разнообразие, который вычисляется путем суммирования значений в строках матрицы расстояний **D**, учитывающей филогенетические различия между видами.

Другой важной составляющей современной экологической парадигмы является функциональный подход, рассматривающий сообщества с точки зрения их отличий по продуктивности, стабильности, скорости усвоения питательных веществ, резистентности к инвазиям, широте спектра реакций отдельных видов на воздействие факторов среды и т.д. *Функциональное разнообразие*, под которым понимают “степень полной функциональной изменчивости (или различий) видов в сообществе” (Tilman, 2001), связано с параметрами ширины и перекрытия пространства экологической ниши, занимаемой каждым видом сообщества. Основные выражения для мер функционального разнообразия основаны на использовании вектора характерных значений (trait values) потенциально важных функциональных признаков, оцениваемых для каждого *i*-го вида из *S* расчетным путем или по эмпирическим данным.

Масон с соавторами (Mason at al., 2005) считают, что функциональное разнообразие не может быть представлено в итоге "одним числом" и выделяют три типа индексов, отражающих соответственно функциональное богатство, функциональную выравненность и функциональную дивергенцию. Например, индекс функциональной дивергенции FD_{var} отражает полную вариацию характерных значений x_i для каждого *i*-го вида относительно среднего значения функционального признака для всего сообщества. Функциональная характерность всего сообщества по каждому показателю или *CWM* (community-weighted mean trait value – Garnier at al., 2004) оценивается как сумма $x_i \cdot p_i$ для всех видов с учетом их относительного обилия p_i :

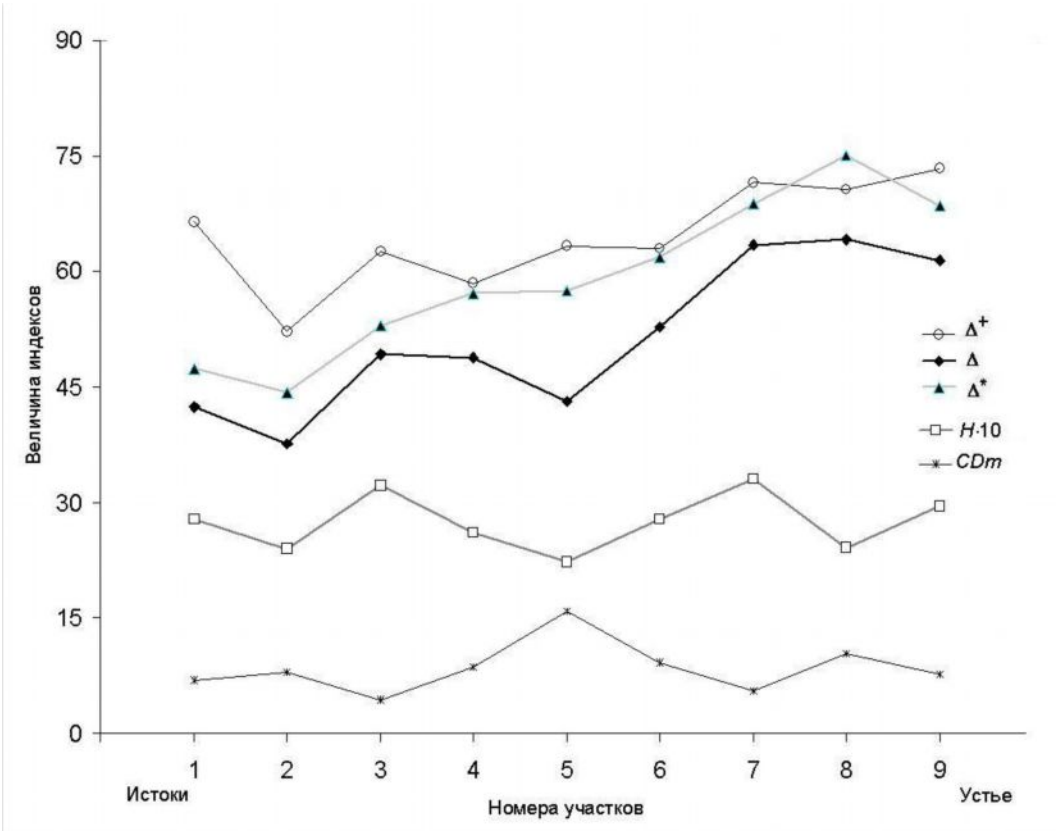
$$FD_{\mathrm{var}} = \frac{2}{\pi}\arctan(5V); \quad V = \sum_{i=1}^{S} p_i (\ln x_i - \sum_{i=1}^{S} p_i \ln x_i)^2 \;; CWM = \bar{x} = \sum_{i=1}^{S} p_i x_i \tag{2.3}$$

Оценка функционального разнообразия также может быть основана на расчете квадратичной энтропии Рао (2.1). Экологические расстояния d_{ij} матрицы дистанций **D** на шкале одного характерного признака легко найти, например, построив функции распределения вероятности признака для обоих сравниваемых видов и вычислить относительную площадь наложения (overlag) двух гауссиан (Lepš at al., 2006).

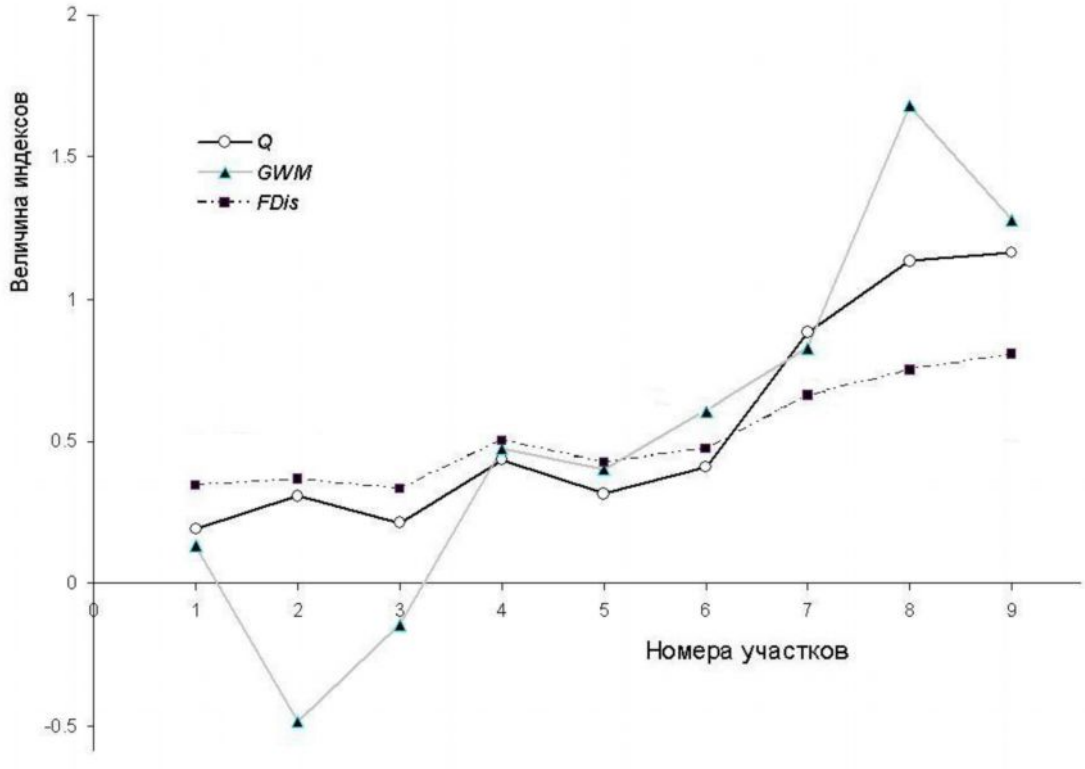
Рассмотрим изменение таксономического и функционального разнообразия макрозообентоса по продольному профилю р. Сок – рис. 2.18. Всего по результатам 97 гидробиологических проб, сделанных на 9 станциях наблюдения, было обнаружено 277 видов и таксономических групп донных организмов. Для каждого таксона из 277 было сделано систематическое описание по 11 классификационным уровням: Species → Genus → Tribe → SubFamily → Family → SubOrder → Order → SubClass → Class → SubPhylum → Phylum. В качестве характерного признака функционального разнообразия использовался логарифм индивидуальной массы особей макрозообентоса.

Результаты статистического анализа, подробно представленные в нашей работе (Шитиков, Зинченко, 2013), позволяют сделать следующие выводы:

° таксономическое своеобразие донных сообществ, оцениваемое по индексу Δ^+, в нижнем участке реки увеличивается, что является свидетельством "вложенности" сообществ (nestedness-эффекта), т.е. биоценозы, расположенные ниже по течению, претерпевают влияние со стороны вышерасположенных композиций видов и "наследуют" часть их видового богатства;



а)



б)

Рис. 2.18. Динамика изменения индексов таксономического (а) и функционального (б) разнообразия сообществ макрозообентоса в направлении от истоков к устью р. Сок;
(а): Δ^+ – таксономическое своеобразие, Δ – таксономическое разнообразие, Δ^* – вариация таксономического своеобразия, *H* - индекс разнообразия Шеннона, $\cdot CD_m$ – индекс доминирования Симпсона *C* с учетом среднего таксономического расстояния D_m;
(б): *Q* - квадратичная энтропия Рао; CWM - средневзвешенная для сообщества величина индивидуальной массы особей; FD_{var} - индекс функциональной дивергенции

° вариация индекса разнообразия Δ имеет не столь выраженную закономерность, поскольку тенденции роста таксономического дерева становятся не столь заметными на фоне более мощной флуктуирующей изменчивости численности видов-доминантов;

° важнейшим фактором продольной биотической изменчивости водотоков является коренная перестройка функциональной роли сообществ: резко возрастает средняя индивидуальная масса особей (*CWM*) и связанная с ней интенсивность оборота органических питательных веществ и биопродукционных процессов в целом, которая достигает максимума на участки 8-9, где донное сообщество в стабилизировано;

° на фоне этих выявленных изменений традиционные индексы видового разнообразия Симпсона и Шеннона, а также таксономической энтропии (2.2) какой-либо отчетливой тенденции для изученного водотока вообще не обнаруживают: их изменчивость носила статистически незначимый флуктуирующий характер, обусловленный частными сменами доминирующих видов.

Сравнение значений индексов для двух или нескольких обследованных биотопов связано с проверкой нулевой гипотезы об отсутствии различий между ними. Для этого разработана рандомизационная процедура (Clarke, Warwick, 1998), в ходе которой многократно (не менее 1000 раз) для каждого участка с исходным видовым богатством *S* генерируются частные псевдовыборки, являющиеся случайной комбинацией *S* видов, извлеченных из общего видового списка видов $S_{общ}$, обнаруженных во всех выборках ($S_{общ} > S$). На основе этих итераций восстанавливается неизвестное распределение значений индекса и находятся оценки его статистических характеристик.

Для последовательности участков р. Сок была построена "туннельная" диаграмма (рис. 2.19), на которой представлены рандомизированное среднее Δ^+_m, кривые доверительных интервалов для 95%-й вероятности в зависимости от видового богатства *S* и облако рассеяния точек эмпирических значений Δ^+ для отдельных биотопов. За пределами нижней доверительной границы Cl_{95} расположились индексы своеобразия для участков 2-4, что указывает на статистически значимое увеличение таксономического разнообразия биоценозов по продольному градиенту водотока.

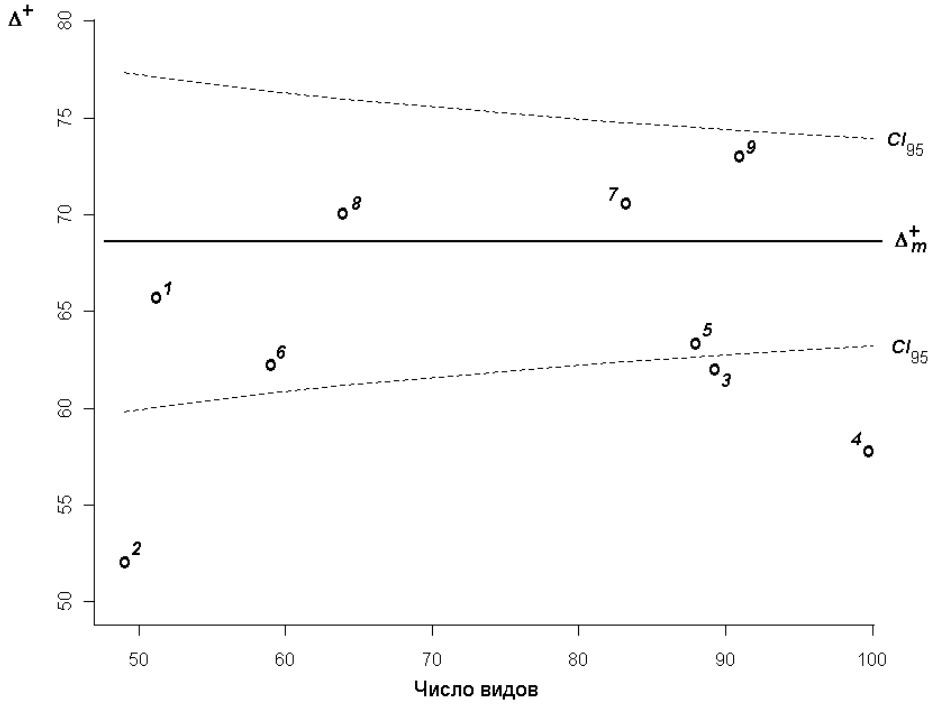


Рис. 2.19. "Туннельная" диаграмма индекса таксономического своеобразия Δ^+ в зависимости от числа видов на отдельных участках р. Сок: Δ^+_m – среднее значения индекса; CI^+_{95} и CI^-_{95} – верхний и нижний пределы доверительного интервала

Для получения более определенного статистического вывода об изменчивости таксономического и функционального разнообразия по профилю реки объединим данные гидробиологических проб на станциях 1-5 и 6-9, и выполним сравнение верхнего и нижнего участков по четырем основным нуль-моделям рандомизации, как это делалось в разделе 2.8 при оценке видового разнообразия.

Таблица 2.6. Сравнение двух участков р. Сок с использованием индекса среднего таксономического своеобразия (Δ^+), таксономического разнообразия (Δ) и средневзвешенной функциональной характерности (CWM); обозначения нуль-моделей приведены в разделе 2.8

Эмпирическая разность $d_{obs}=(x_2 - x_1)$	Тип модели	95% доверительные интервалы d_{ran} при справедливости H_0	Вероятность $Pr(d_{ran} \geq d_{obs})$
$\Delta^+_2 - \Delta^+_1 =$ 71.6 -60.6 = 11.0	1. Перестановочная модель EE	-4.16 ÷ 4.01	**0.001**
	2. Обмен в строках (модель EF1)	-5.44 ÷ 5.88	**0.001**
	3. Фиксированная по видам EF2	-5.51 ÷ 5.56	**0.001**
	4. Дважды фиксированная модель FF	-0.43 ÷ 0.41	**0.001**
$\Delta_2 - \Delta_1 =$ 56.42 - 55.36 = 1.06	1. Перестановочная модель EE	-22.2 ÷ 23.2	0.916
	2. Обмен в строках (модель EF1)	-11.9 ÷ 12.6	0.865
	3. Фиксированная по видам EF2	-7.89 ÷ 8.34	0.802
	4. Дважды фиксированная модель FF	-0.189 ÷ 0.185	**0.001**
$CWM_2 -$ $CWM_1 =$ 0.6 - (-0.23) = 0.83	1. Перестановочная модель EE	-2.46 ÷ 2.64	0.462
	2. Обмен в строках (модель EF1)	-0.74 ÷ 0.76	**0.023**
	3. Фиксированная по видам EF2	-0.58 ÷ 0.60	**0.002**
	4. Дважды фиксированная модель FF	-0.01 ÷ 0.01	**0.001**

Результаты расчетов показывают, что верхний и нижний участки р. Сок статистически значимо отличаются по своему таксономическому своеобразию Δ^+ и средней массе тела бентосных животных *CWM* (с использованием наиболее экологически обоснованных нуль-моделей). Однако предположение о различии таксономического разнообразия с учетом численности видов по индексу Δ своего подтверждения не получило.

К разделу 2.9:

```
#  Загрузка исходных данных для расчета из листов файла xls
library(xlsReadWrite)
# На 1-м листе  - значения численностей по видам, на 2-м листе  - систематика видов
TTB <- t(read.xls("Сок_таксон.xls", sheet = 1, rowNames=TRUE))
TTB.taxon <- read.xls("Сок_Таксон.xls", sheet = 2, rowNames=TRUE)
#  На  3-м листе  - значения индивидуальной массы по видам
TR  <-  read.xls("Сок_таксон.xls", sheet = 3, rowNames=TRUE)
# Свертывание данных до двух участков р. Сок и сравнение показателей
s1 <- apply(TTB[1:5,],2,sum) ; s2 <- apply(TTB[6:9,],2,sum) ; TT2S <- t(data.frame(s1, s2))
library(vegan)              # Включение пакета vegan
# Определение функции оценки p-значения при сравнении двух участков
# data - таблица данных, Nperm - число итераций, fixedmar, shuffle - параметры permatfull
# divdiff - внешняя функция расчета разности индексов разнообразия
p.rand <- function(data, fixedmar, shuffle, Nperm=999) {
       d_emp <- divdiff(data)  ; d_rand <- as.numeric(rep(NA,Nperm))
       for (i in 1:Nperm) {
              x3 <- permatfull(data, fixedmar = fixedmar,shuffle = shuffle, times = 1)
              TT_rand <- as.matrix(x3$perm[[1]]) ; d_rand[i] <- divdiff(TT_rand) }
       RandRes (d_emp, d_rand , Nperm) }
source("print_rezult.r")  # Загрузка функций вывода результатов

#    ------------    a) Таксономическое разнообразие
# Формирование матрицы таксономических дистанций между всеми видами
```

```
taxdisB <- taxa2dist(TTB.taxon, varstep=TRUE)
# Расчет индексов таксономического разнообразия по 9 станциям р. Сок (вывод в буфер обмена)
 mod <- taxondive(TTB, taxdisB); write.table(summary(mod), file="clipboard", sep="\t")
plot(mod)  #  Вывод туннельной диаграммы
#  Расчет таксономической энтропии Шеннона
TaxEnt <- function (spec, taxdis) {
       k_sum <- apply(as.matrix(taxdis),1,sum); k_sum <- k_sum/sum(k_sum)
       SH <- as.numeric(rep(NA, nrow(spec)))
       for (i in 1:nrow(spec)) SH[i] <- sum(spec[i,]*log(k_sum))/sum(spec[i,])
       return (SH) }
TaxEnt(TTB, taxdisB)
# Расчет таксономических индексов для 2 участков
 mod2 <- taxondive(TT2S, taxdisB);  summary(mod2) ; TaxEnt(TT2S, taxdisB)
#       Рандомизационный тест сравнения индексов таксономического разнообразия
#  Функция сравнения индекса Дельта+  для двух участков
divdiff <- function(x) { m <- taxondive(x, taxdisB); m$Dplus[2]-m$Dplus[1]}
#    Аналогичная функция для сравнения индекса Дельта
#   divdiff <- function(x) { m <- taxondive(x, taxdisB); m$D[2]-m$D[1]}
p.rand (TT2S,fixedmar = "none", shuffle = "samp")     # 1. Перестановочная модель EE
p.rand (TT2S,fixedmar = "columns", shuffle = "samp")  # 2. Обмен в строках (модель EF1)
p.rand (TT2S,fixedmar = "columns", shuffle = "both")  # 3. Фиксированная по видам модель EF2
p.rand (TT2S,fixedmar = "both", shuffle = "both")     # 4. Дважды фиксированная модель FF

#    -------------    в) Функциональное разнообразие
#   Оценка распределения характеристики и логарифмирование значений
hist(TR$IM) ; TR$IMlog <- log2(TR$IM) ; hist(TR$IMlog)
# Применение логарифмической трансформации к численностям
bentosTraits <- subset(TR, select = "IMlog")  ;  bentosAbun   <-  decostand( TTB, "log")
library(FD)    #   Загрузка пакета и выполнение расчета функционального разнообразия
result <- dbFD(bentosTraits, bentosAbun)
 write.xls(as.data.frame(result), "fdr.xls") #   Результаты экспортированы в файл Excel
# Расчет функциональных индексов для 2 участков
bentosAbun2s   <-  decostand(TT2S , "log") ; result2s <- dbFD(bentosTraits, bentosAbun2s)
#  Функция permatfull может работать только с целочисленными значениями имитируемой матрицы
CW <- functcomp(bentosTraits, TT2S)  # Сравнение проводим только по   CWM
divdiff <- function(x) { colnames(x) <- rownames(bentosTraits)
                            CW.val <- functcomp(bentosTraits, x); CW.val[2,] - CW.val[1,]}
p.rand (TT2S,fixedmar = "none", shuffle = "samp")     # 1. Перестановочная модель EE
p.rand (TT2S,fixedmar = "columns", shuffle = "samp")  # 2. Обмен в строках (модель EF1)
p.rand (TT2S,fixedmar = "columns", shuffle = "both")  # 3. Фиксированная по видам модель EF2
p.rand (TT2S,fixedmar = "both", shuffle = "both")     # 4. Дважды фиксированная модель FF
```

3. СТАТИСТИЧЕСКИЕ ЗАВИСИМОСТИ И СВЯЗИ МЕЖДУ ПЕРЕМЕННЫМИ

3.1. Оценка парной корреляции с использованием рандомизации

Важным компонентом статистического анализа является оценка степени корреляционной связи между отдельными переменными. Пусть мы имеем совместное нормальное распределение двух случайных величин (*X, Y*), плотность вероятности которого определяется пятью моментами: средними μ_X , μ_Y, дисперсиями σ_X , σ_Y и *коэффициентом корреляции* $\rho_{XY} = E[\frac{(X-\mu_X)(Y-\mu_Y)}{\sigma_X\sigma_Y}]$, где E – символ математического ожидания. Если получена случайная выборка из генеральной совокупности объемом *n* и для каждого ее *i*-го элемента определены сопряженные значения реализаций случайных величин (x_i, y_i), то количественной оценкой ρ является коэффициент парной корреляции Пирсона: $R_{xy} = \frac{Cov(x,y)}{\sqrt{S_x^2}\sqrt{S_y^2}}$, где $S_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2$ и $S_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2$ – выборочные дисперсии значений *X* и *Y*, а $Cov(x,y) = \frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})$ – ковариация, характеризующая совместное распределение (*X, Y*) в двумерном евклидовом пространстве.

Нулевая гипотеза H_0: $\rho_{XY} = 0$ об отсутствии линейной связи утверждает, что анализируемые переменные независимы между собой настолько, что все возможные сочетания реализации пар величин x_i и y_i равновероятны, а те или иные закономерности в их совместной изменчивости объясняются лишь случайными механизмами порождения данных. Если оценка *R* статистически незначима (т.е. слишком близка к 0), то можно сделать предположение, что показатели не зависят друг от друга, либо эта зависимость носит отчетливо нелинейный характер. Анализу должна предшествовать проверка нормальности распределения обоих вариационных рядов (*X, Y*): если это допущение отклоняется, то рекомендуется отдавать предпочтение непараметрическим критериям.

Параметрический подход к оценке значимости коэффициентов корреляции основан на аппроксимации некоторых статистик от *R* теоретическими распределениями *t* Стьюдента или *Z* Фишера. Если, например, предположить, что наблюдения в выборке независимы и нормально распределены, то отношение $t = R\sqrt{n-2}/\sqrt{1-R^2}$ проверяется с использованием распределения Стьюдента с $\nu = n - 2$ степенями свободы. Аналогично тесту для связанных выборок (раздел 2.4), полученное при этом значение вероятности *p* является оценкой статистической значимости α нулевой гипотезы о равенстве нулю коэффициента корреляции H_0: $\rho = 0$.

Если для двумерной корреляционной модели параметр ρ оказался значим, то имеет смысл найти для него интервальную оценку (т.е. построить доверительный интервал). Плотность распределения выборочного коэффициента корреляции имеет сложный вид, поэтому используют специальные аппроксимирующие процедуры, такие, как *Z*-трансформация Фишера, называемая также преобразованием обратного гиперболического тангенса. Случайная величина $Z = 0.5\ln[(1+\rho)/(1-\rho)]$ при $n > 10$ распределена по нормальному закону $N(\mu_Z, \sigma_Z^2)$, $\sigma_Z^2 = 1/(n-3)$, и при неавтоматизированном варианте расчетов представлена в таблицах. Тогда алгоритм нахождения доверительных интервалов ρ сводится к следующему:

- по таблице *Z*-трансформации Фишера находят значение z_R, соответствующее выборочному коэффициенту корреляции *R*;
- строят интервальную оценку для матожидания *Z*: $z_R - t_\alpha \sigma_Z \le E(Z) \le z_R + t_\alpha \sigma_Z$;
- граничные значения z_{min} и z_{max} доверительного интервала E(*Z*) при доверительной вероятности $\gamma = 1 - 2\alpha$ с помощью тех же таблиц *Z*-трансформации пересчитывают в граничные значения для ρ: $R_{min} \le \rho \le R_{max}$.

Непараметрические коэффициенты ранговой корреляции Спирмена и Кендалла являются менее мощными, т.к. оценивают уже не параметры совместного двумерного распределения случайных величин (X, Y), а некоторую меру равновероятности рангов векторов данных. В то же время, они позволяют выявлять корреляцию при нелинейной связи между переменными, даже если эта зависимость носит немонотонный характер.

Предположим, что р. Сок [пример П2] на всем ее протяжении от истоков до устья разбита на 13 участков, и мы хотим проанализировать изменчивость видового состава. Рассчитаем коэффициент корреляции между значениями температуры воды t, которая в данном случае олицетворяет продольный градиент реки, и долей d_{DO} донных организмов Diamesinae+Orthocladiinae[6] в общей численности зообентоса.

Рандомизационная процедура для оценки линейной связи двух переменных сводится к тому, что многократно выполняются случайные перестановки значений одной переменной относительно другой (например, температура для участка 1 ставится в соответствие с долей диамезин для участка 7 и далее в таком же перетасованном беспорядке). Для каждой имитируемой выборки рассчитывается рандомизированный коэффициент корреляции R_{ran}, математическое ожидание которого равно 0, поскольку каждый x_i *случайно* связан со значениями y_i и зависимость между переменными разрушена.

Выполнив B = 5000 таких итераций, получим гистограмму распределения моделируемой статистики при справедливости нулевой гипотезы и подсчитаем количество случаев, когда коэффициент корреляции R_{ran} для рандомизированных комбинаций превысил по абсолютной величине коэффициент корреляции R_{obs} = -0.678 для эмпирических наблюдений. На рис. 3.1 таких областей две: слева с частотой случаев b_1 = 48 при $R_{ran} < R_{obs}$ и справа с частотой b_2 = 33 при $R_{ran} > |R_{obs}|$; $b = b_1 + b_2 = 48 + 33 = 81$.

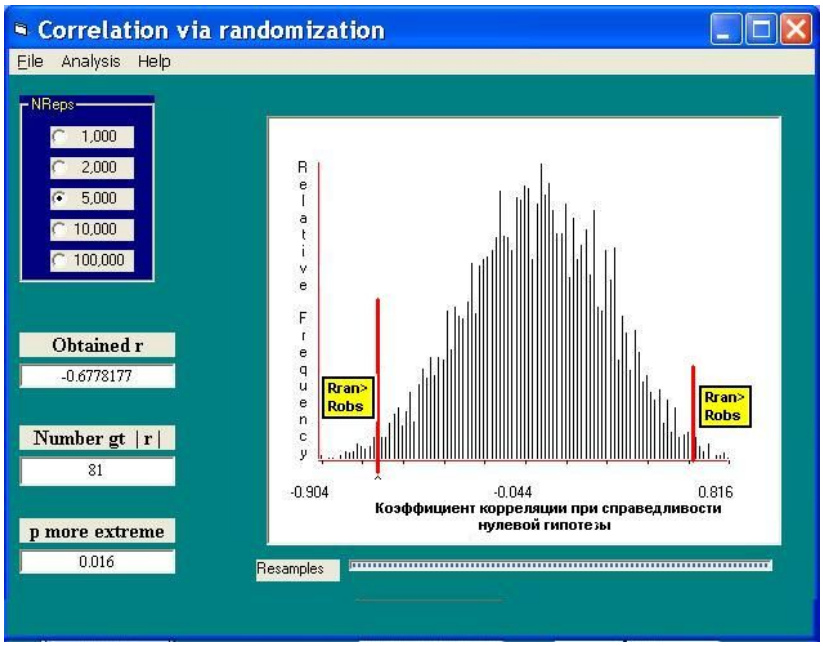


Рис. 3.1. Распределение коэффициента корреляции R между обилием Diamesinae+Orthocladiinae и температурой воды при нулевой гипотезе. Obtained R – значение R для исходных данных; Number gt |r| и p more extreme – соответственно частота и вероятность превышения этого значения для рандомизированных данных

Частоты b_1 и b_2 позволяют рассчитать p-значения при различной формулировке альтернативной гипотезы. Если есть намерение проверить одностороннюю гипотезу H_1: $R_{obs} < 0$, предполагающую обратную связь между переменными, то $p = (b_1 + 1)/(B + 1) = 49/5001 = 0.0096$. По умолчанию формулируется двусторонняя альтернатива H_1: $R_{obs} \neq 0$ (связь между переменными имеет место) и тогда $p = (b + 1)/(B + 1) = 81/5001 = 0.016$. В

[6] Диамезины и ортокладеины: семейства видов мотыля - личинок хирономид или комара-звонца.

любом варианте тестирования увеличение доли диамезин по мере уменьшения температуры воды можно считать обоснованным.

Заметим также, что различия *p*-значений, полученных параметрическим ($p_{\text{парам}}$) и рандомизационным ($p_{\text{ранд}}$) методами для всех коэффициентов корреляции, не слишком велики:

- коэффициент корреляции Пирсона $R = -0.68$ ($p_{\text{парам}} = 0.011$, $p_{\text{ранд}} = 0.016$);
- коэффициент ранговой корреляции Спирмена $r_S = -0.564$ ($p_{\text{парам}} = 0.0445$, $p_{\text{ранд}} = 0.053$);
- коэффициент ранговой корреляции Кендалла $\tau = -0.426$ ($p_{\text{парам}} = 0.0427$, $p_{\text{ранд}} = 0.050$).

Если многократно формировать *n*-мерные случайные повторные выборки по алгоритму "выбора с возвращением" из номеров строк исходного набора данных (x_i , y_i), то можно получить статистическое распределение коэффициента корреляции и рассчитать бутстрепированные доверительные границы для ρ. Для нашего примера интервальные оценки, полученные методом процентилей ($-0.95 \le \rho \le -0.19$), оказались несколько шире 95%-го доверительного интервала, построенного с использованием *Z*-трансформации Фишера ($-0.895 \le \rho \le -0.2$).

К разделу 3.1:

```
#   Оценка значимости коэффициента корреляции
#   Определяем исходные данные:  NDO - доля численности Diamesinae+Orthocladiinae
#    T  -  температура воды при взятии гидробиологической пробы
NDO <- c(88.9,94.9,70.8,46.4,31.0,66.5,83.6,71.9,59.6,22.5,29.2,6.5,17.5)
T <- c(14.8,14.5,11.5,12.6,12,14.5,13.3,16.7,16.9,20.6,21.2,22.5,22.4)
#  Рассчитываем p-значения коэффициентов корреляции по асимптотическим формулам
(cors <- cor.test(NDO,T,method = "pearson"))
library(psych) # Используем функции Z-трансформации Фишера из другого пакета
zs <- fisherz(cors$estimate); rs <- fisherz2r(cors$estimate) ; round(zs,2)
r.con(cors$estimate,length(T))
cor.test(NDO,T,method = "kendall");cor.test(NDO,T,method = "spearman")
source("print_rezult.r")  # Загрузка функций вывода результатов
#   Функция рандомизационного теста значимости корреляции
PermCor <- function (X, Y, method="pearson" , Nperm=5000) {
  PermArray <- as.numeric(rep(NA, Nperm)) ;  for (i in 1:Nperm) #  случайно перемешиваем Y
  PermArray[i] <- as.numeric(cor.test(X,sample(Y,length(Y)),method = method)$estimate)
  return (RandRes(cor.test(X,Y,,method = method)$estimate, PermArray, Nperm)) }
#   Функция оценки доверительных интервалов коэффициента корреляции
BootCor <- function (X, Y, method="pearson" , Nboot=5000) {
  Remp <- as.numeric(cor.test(X,Y,method = method)$estimate) ; n <- length(Y)
  BootArray <- as.numeric(rep(NA, Nboot)) ; for (i in 1:Nboot) {
  Ind <- sample.int(n, size = n, replace = TRUE)  # о выбираем индексы строк с возвращением
  BootArray[i] <- as.numeric(cor.test(X[Ind],Y[Ind],method = method)$estimate) }
  return (BootRes(BootArray, Remp)) }
#   Выполнение расчетов
PermCor(NDO,T) ; BootCor(NDO,T)
PermCor(NDO,T, method = "kendall") ; BootCor(NDO,T, method = "kendall")
PermCor(NDO,T, method = "spearman") ; BootCor(NDO,T, method = "spearman")
```

3.2. Анализ связи между признаками в таблицах сопряженности

Часто объекты изучаемой выборки описаны категориальными переменными. Тип данных, которыми представлены такие показатели, носит вполне определенный характер: они измеряются в шкалах наименований независимых классов, порядковых (ординальных) шкалах, либо сведены к таковым в ходе предварительной обработки.

Предположим, что признак *A* имеет *r* градаций (или уровней) A_1, A_2, …, A_r, а признак *B* подразделяется на *c* градаций B_1, B_2, …, B_c. Тогда в "свернутом" виде результаты наблюдений можно представить *таблицей сопряженности* $r \times c$, в ячейках которой проставлены частоты событий n_{ij}, т.е. количество объектов выборки из *n* элементов, обладающих комбинацией уровней A_i и B_j. Маргинальные значения таблицы

$n_{i\bullet}$ и $n_{\bullet j}$ являются суммами частот n_{ij} по столбцам и строкам соответственно (замена индекса точкой означает результат суммирования по этому индексу).

Если между переменными *A* и *B* имеется взаимно однозначная прямая или обратная функциональная связь, то все частоты n_{ij} концентрируются по одной из диагоналей таблицы. При связи, не столь сильной, некоторое число наблюдений попадает и на недиагональные элементы. Если признаки *A* и *B* будут независимыми, то значение, принятое признаком *A,* никак не влияет на вероятности возможных значений признака *B*:

$$P(B_j/A_i) = P(B_i) \text{ или } P(A_i,B_j) = P(A_i)\,P(B_j)$$

Значения использованных вероятностей нам неизвестны, однако по теореме Бернулли при большом объеме выборки ($n \to \infty$) значения в ячейках таблицы сопряженности будут являться оценками этих вероятностей: и тогда соответствующие величины *p* можно трактовать как ожидаемые частоты:

$$\frac{n_{ij}}{n} \to p_{ij}; \quad \frac{n_{i\bullet}}{n} \to p_{i\bullet}; \quad \frac{n_{\bullet j}}{n} \to p_{\bullet j}; \quad n = \sum_{i=1}^{r}\sum_{j=1}^{c} n_{ij}\ .$$

Поскольку при независимости признаков справедливо $p_{ij} = p_{i.} \cdot p_{.j}$, то проверка нулевой гипотезы сводится к оценке, насколько близки значения фактических и ожидаемых частот, т.е. $n_{ij} \approx \frac{n_{i\bullet}n_{\bullet j}}{n}$.

Методы сравнения эмпирических (*E*) и теоретических (*T*) частот по А.Брандту и Г.Снедекору (Brandt, Snedecor) могут основываться на расчете критерия согласия χ^2, оценивающего меру близости по всем ячейкам таблицы сопряженности:

$$\chi^2 = \sum \frac{(E-T)^2}{T} = \sum_{i=1}^{r}\sum_{j=1}^{c} \frac{(n_{ij} - \frac{n_{i\bullet}n_{\bullet j}}{n})^2}{\frac{n_{i\bullet}n_{\bullet j}}{n}}.$$

Согласно теоремы Пирсона-Фишера для независимых признаков при неограниченном росте числа наблюдений распределение случайной величины χ^2 стремится к распределению “хи-квадрат”, поэтому нулевую гипотезу о независимости можно принять, если эмпирическое значение χ^2 не превосходит критической величины с $df = (r-1)(c-1)$ степенями свободы для заданного уровня значимости.

Другой вариант проверки нулевой гипотезы связан с рандомизацией и при этом результат будет базироваться на свойствах полученной выборочной совокупности, а не на предположениях о форме распределения или математических свойствах аппроксимаций. Определим *нуль-модель* **FF** (Patefield, 1989; Gotelli, Entsminger, 2003) с фиксированными суммами строк и столбцов как множество *B* таблиц сопряженности со случайными значениями частот n_{ij}^*, но постоянными маргинальными значениями $n_{i\bullet}$ и $n_{\bullet j}$, такими же, как и в исходной таблице наблюдений. Если выбрать формулу для произвольного критерия *G*, наиболее подходящего для решения поставленной задачи, то оценка *p*-значения сводится к следующей знакомой процедуре:

- рассчитывается наблюдаемое для реальных выборок значение статистики G_{obs};
- для каждой *и*-й реализации нуль-модели **FF** из *B* рассчитывается рандомизированное значение статистики G_{ran} ;
- подсчитывается число *b* случаев, когда значения G_{ran} превысили[7] G_{obs}., и *p*-значение определяется как доля $p = (b+1)(B+1)$.

Кроме χ^2 в качестве конкретных дефиниций критерия *G* могут быть использованы различные варианты статистик однородности частот (Кресси-Рида, Хеллингера,

[7] В случае, если высокое значение статистики свидетельствует о больших различиях между выборками. Если же, наоборот, низкие значения статистики свидетельствуют о различиях, то подсчитывается доля G_{ran}, меньших G_{obs}.

Зелтермана), максимум остатков Пирсона или любая из многочисленных мер сопряженности признаков. Мы здесь не касаемся "тонкой материи" выбора, какая из этих формул лучше подойдет исследователю в конкретных условиях для решения определенной задачи: эти проблемы обсуждаются, например, в справочнике Гайдышева (2001). Однако рассмотрим на примерах программную реализацию в среде R двух специальных подходов к анализу данных экологического мониторинга, описанных нами ранее (Шитиков и др., 2008), и сравним их с традиционными методами анализа. Все расчеты выполнены в статистической среде R по скриптам, приведенным в конце раздела.

В четырех физико-географических районах Крыма были изучены локальные популяции улиток *Helix albescens* [пример П7], у которых отмечен полиморфизм по характеру опоясанности раковины. В лабораторных условиях было проанализировано 3115 моллюсков и были рассчитаны частоты отдельных морф – см. табл. 3.1. Ставится задача выяснить степень фенетической дифференциации между отдельными группами популяций моллюска, взятых из различных регионов.

Таблица 3.1. Численность различных морф по характеру опоясанности раковины моллюска *H. albescens* из различных популяций Крыма

Морфа *A* / регион *B*	Симферо-польская	Южная	Степная	Керченская	Итого
12345	270	448	639	173	1530
1(23)45	83	126	369	342	920
10345	15	69	58	6	148
12045	83	18	79	5	185
12305	2	15	26	6	49
10045	36	5	36	0	77
12005	1	22	23	0	46
10005	24	34	78	1	137
10305	4	48	37	1	90
02045	1	2	0	0	3
00005	3	0	0	0	3
00000	1	17	0	0	18
1(23)(45)	0	0	4	3	7
123(45)	1	0	0	1	2
Итого	524	804	1349	538	3215
Энтропия *H*	2.1470	2.1856	2.2116	1.2416	2.2014

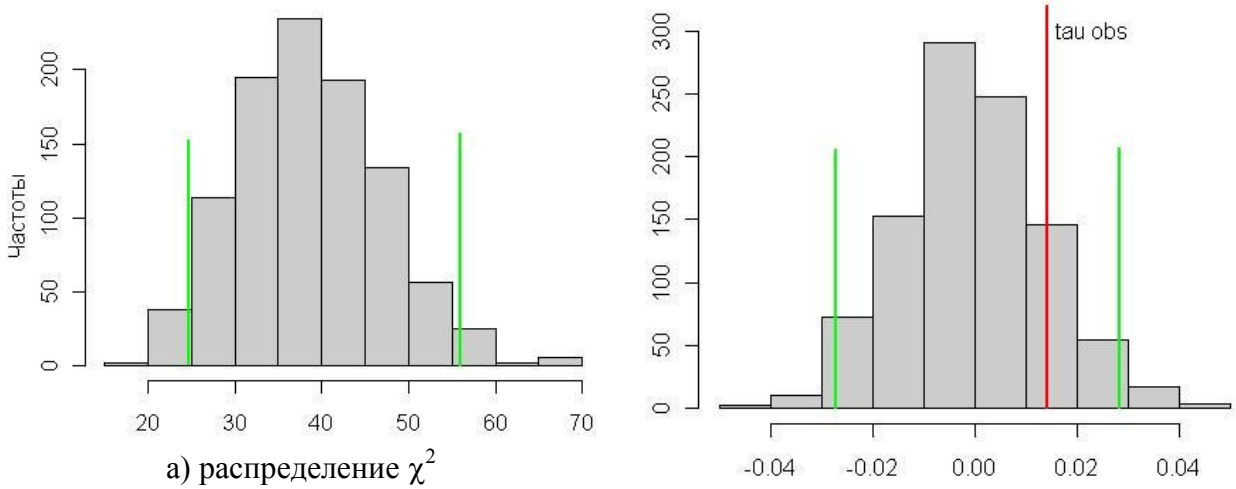
Поскольку в некоторых ячейках таблицы сопряженности значения частот оказались менее 5, использование критерия χ^2 и оценка его *p*-значения с использованием теоретического распределения “хи-квадрат” не является корректным (в табл. 3.2 эти величины приведены лишь для сравнения). Однако, если абстрагироваться от асимптотических предположений и интерпретировать χ^2 как некую произвольную статистику, тестирующую однородность частот на основе рандомизационной процедуры, то можно предположить, что ограничения на величину частот уже не являются столь категоричными. Тогда из 6 декларированных условий применения критерия χ^2 остаются только требования к независимости формирования групп и выборки самих наблюдений (Good, 2005б). Значение χ^2_{obs} на рис. 3.2а находится значительно правее распределения χ^2_{ran} при справедливости H_o, а *p*-значение, полученное рандомизацией, намного меньше 0.05, и поэтому у нас есть все основания отклонить нулевую гипотезу об отсутствии региональной изменчивости популяций моллюска по частотам встречаемости разных вариантов формы раковины.

Очевидно, что к аналогичному выводу приведет использование других мер сопряженности, основанных на χ^2 (коэффициенты Пирсона, Крамера, Чупрова и проч.), поскольку при любом монотонном преобразовании критерия χ^2 предупорядоченность его

значений и, следовательно, результаты рандомизации не меняются.

Таблица 3.2. Результаты оценки независимости и сопряженности признаков регионально-морфологической изменчивости моллюска *H. albescens* с использованием различных критериев (*A* - морфа, *B* - регион)

Использованные критерии		Эмпирическое значение G_{obs}	*p*-значение асимптотич.	Рандомизация (1000 итераций)	
				Доверительные интервалы (95%)	*p*-значение $G_{ran} > G_{obs}$
Статистика χ^2		777.6	$2.2 \cdot 10^{-16}$	24.22 ÷ 56.6	0.001
τ Валлиса-Гудмена-Краскела	*A* / *B*	0.0663	$3 \cdot 10^{-12}$	0.00237 ÷ 0.00539	0.001
	B / *A*	0.055	$2 \cdot 10^{-13}$	0.000349 ÷ 0.0016	0.001
λ Гуттмена-Гудмена-Краскела	*A* / *B*	0.1002		0	0.001
	B / *A*	-0.0783		-0.107 ÷ -0.103	0.001
τ_c Кендалла		0.0146		-0.0275 ÷ 0.0277	0.287



а) распределение χ^2 (эмпирическое значение равно 777.6)

б) распределение τ_c Кендалла

Рис. 3.2. Распределение статистик независимости и связи для таблицы сопряженности частот встречаемости экоморф моллюска *H. albescens* при нулевой гипотезе (зелеными линиями показаны доверительные интервалы)

Оценка риска отклонить нулевую гипотезу о независимости переменных – не самоцель. Установив наличие связи, необходимо измерить ее силу с тем, чтобы иметь возможность сравнивать компоненты системы и выделять информативные признаки. Наиболее теоретически проработанными считается (Елисеева, Рукавишников, 1977) семейство несимметричных мер Гудмена-Краскела (Goodman, Kruscal, 1954), использующих принцип “пропорциональной предикции”: чем меньше относительная вероятность ошибки предсказания значения зависимого признака по значению независимого, тем больше сила связи.

Мера τ Гудмена-Краскела (в русскоязычной литературе – коэффициент Валлиса) вычисляется по формуле $\tau_{Y/X} = \left[\sum_{i=1}^{r} \sum_{j=1}^{c} (p_{ij} - p_{i\bullet} p_{\bullet j})^2 / p_{i\bullet} \right] / (1 - \sum_{j=1}^{c} p_{\bullet j}^{\,2})$. Ее интерпретация чрезвычайно проста: если, например, $\tau_{y/x} = 0.50$, то знание *X* уменьшает число ошибок прогноза *Y* вдвое. Другая мера λ Гудмена-Краскела (впервые описанная Л. Гуттменом в 1941 г.) также отражает редукцию ошибок предсказания. Его формула приспособлена для быстрых вычислений, но имеет серьезный недостаток: при определенных комбинациях маргинальных категорий он обращается в нуль (это имело место и в нашем случае для всех 1000 нуль-модельных значений). По абсолютной величине этих мер судить о силе связи практически невозможно, т.к. диапазон их варьирования сильно зависит от размерности *r*×*s* таблиц сопряженности и степени "перекоса" частот. Однако

использование рандомизации позволяет четко сказать, насколько далеко эмпирическое значение мер от интервальной оценки при справедливости нулевой гипотезы. В частности (см. табл. 3.2), использование мер Гудмена-Краскела свидетельствует о существовании отчетливой зависимости изменчивости морф изучаемой улитки от регионального фактора, которую нельзя объяснить случайными обстоятельствами.

Третью группу мер сопряженности составляют коэффициенты, основанные на рангах, которые позволяют извлечь информацию о направлении связи между признаками, используя понятие коррелируемости на основе подсчета числа пар объектов с взаимно возрастающими, взаимно убывающими и равными значениями признаков. Поскольку они используются лишь, когда входы таблицы сопряжённости упорядочены (т.е. градации признаков измерены в порядковой шкале), их применение в контексте рассматриваемой задачи совершенно некорректно. Однако в методических целях мы выполнили анализ на основе коэффициента τ_c Кендалла, который оптимизирован для анализа несимметричных таблиц, и убедились в том, что $\tau_{c\,obs}$ накрывается доверительным интервалом $\tau_{c\,ran}$, построенным при справедливости нулевой гипотезы (рис. 3.2б). Разумеется, это не является свидетельством отсутствия связи между признаками, а объясняется номинальным характером используемых категорий.

«Рассмотренные выше традиционные меры связи – это сугубо эвристические конструкции, интерпретация и математико-статистическое обоснование которых оставляет желать много лучшего» (Елисеева, Рукавишников, 1977, с. 89). Например, корректное использование статистики χ^2 связано с целым рядом требований: отсутствие нулевых частот и ограничения на минимальную величину n_{ij}, достаточная "насыщенность" таблицы сопряженности и отсутствие ее резкого "перекоса", необходимость введения поправок на непрерывность и т.д.

В последнее время появилось много публикаций, в которых продемонстрированы возможности применения энтропийно-информационного анализа в различных областях биологической науки, физиологии и медицине и др. Поэтому рассмотрим применение в качестве критерия G статистику, основанную на популярном в экологии индексе Шеннона. Ниже описана реализация алгоритма, предложенного С.С.Крамаренко, для сравнения k ($k > 2$) оценок энтропии с разложением суммарной изменчивости комплекса наблюдений на межгрупповую и остаточную компоненты.

При выполнении энтропийного дисперсионного анализа (ЭДА) для каждой включенной в анализ региональной группы популяций, а также для суммарных данных рассчитываются соответствующие оценки энтропии (см. табл. 3.1):

$$H_i = -\sum_{j=1}^{r} \frac{n_{ij}}{n_j} \cdot \log_2 \frac{n_{ij}}{n_j}; \qquad H_T = -\sum_{j=1}^{r} \frac{n_j}{N} \cdot \log_2 \frac{n_j}{N}.$$

Затем необходимо рассчитать следующие величины, соответствующие суммам квадратов ANOVA: $C_T = N \cdot H_T$, $C_R = \sum_{i=1}^{k} N_i \cdot H_i$, $C_A = C_T - C_R$;

и средним квадратам с учетом числа степеней свободы: $MS_A = \frac{C_A}{k-1}$ и $MS_R = \frac{C_R}{N-k}$.

Разумеется, традиционный метод проверки гипотез, используемый в классическом дисперсионном анализе, не может быть использован при сравнении энтропий, поскольку приведенные средние квадраты не являются оценками дисперсий, а их отношение $F_H = MS_A / MS_R$ не может быть аппроксимировано F-распределением Фишера-Снедекора. Поэтому для проверки нуль-гипотезы о равенстве всех оценок энтропии в анализируемых выборках используем показатель, рассчитанный по формуле $\eta = (F_H - 1)/(F_H - 1 + N_0)$, где N_0 – средняя взвешенная численность объектов в группах, $N_0 = \frac{1}{c-1} \cdot (N - \frac{1}{N} \sum_{i=1}^{c} N_i^2)$.

Стандартная таблица дисперсионного анализа, рассчитанная по данным табл. 3.1, имеет вид:

Компоненты дисперсии	Сумма квадратов *C*	Степени свободы *df*	Средние квадраты *MS*	F_H , N_0	η
Между группами *A*	543.9	3	181.3	89.1	0.1042
Внутри групп *R*	6533	3211	2.035	757.5	
Общая *T*	7077	3214	2.202		

Если оценки вероятностей n_{ij} / n_j объектов разного типа равны или пропорциональны во всех сравниваемых выборках, то показатель η будет равен нулю. И, наоборот, при отчетливой блочной или существенно неоднородной структуре распределения частот по столбцам (в нашем случае, морф по группам популяций), а общее число особей *N* достаточно велико, показатель η будет стремиться к единице. Поэтому, как и в общем случае мер корреляции или сопряженности, проверку предположения о равенстве оценок энтропии в сравниваемых выборках можно свести к проверке нулевой гипотезы: $H_0 : \eta = 0$ против альтернативы $H_A : \eta \neq 0$.

Выполнение рандомизационного теста с использованием 1000 реализаций нуль-модели **FF** дает нам 95% доверительные интервалы при справедливости нулевой гипотезы η_{ran} = (0.00258 ÷ 0.00717) , т.е. η_{obs} = 0.1042 статистически значимо превышает эти значения с вероятностью ошибиться *p* = 0.001. Следовательно, мы можем сделать вывод, что разнообразие морфоформ улитки (в смысле их представленности и выравненности вероятности обнаружения) имеет региональную изменчивость.

Второй пример [П8] основан на качественном анализе видовой структуры ихтиоценоза Чебоксарского водохранилища по результатам многолетних наблюдений. Всего использовались данные по 152 съемкам методом неводного лова, в которых всего зафиксировано *r* = 34 вида ихтиофауны. Все 67 станций, где проводился отлов рыбы, были отнесены к *c* = 5 типичным участкам. В табл. 3.3 приведен фрагмент списка видов и показатели их встречаемости: количество проб n_i, в которых обнаружен вид, либо доля t_i, т.е. отношение частоты к общему числу проб.

Таблица 3.3. Распределение встречаемости видов рыб на участках Чебоксарского водохранилища

Вид	Верх-нереч ной	Сред-нереч ной	Озер -ный	При-плотин-ный	Река Ока	Водохранили-ще в целом		Крите рий χ^2	*p*-вероят-ность
						n	*t*		
Бычок кругляк	**0**	**5**	**8**	**8**	**5**	**26**	**17.1**	**20.65**	**0.0005**
Густера	2	15	7	5	2	31	20.4	3.998	0.4122
Елец	9	28	16	11	8	72	47.4	2.059	0.7347
Ерш	8	15	8	5	7	43	28.3	2.549	0.6482
Жерех	**3**	**26**	**11**	**10**	**9**	**59**	**38.8**	**10.77**	**0.0286**
Лещ	18	45	23	8	14	108	71.0	8.081	0.0934
Окунь	20	48	26	15	14	123	80.9	0.927	0.9286
Плотва	19	52	26	17	15	129	84.9	2.483	0.6637
Тюлька	4	15	5	1	6	31	20.4	6.114	0.1969
Уклея	**19**	**32**	**15**	**9**	**12**	**87**	**57.2**	**10.11**	**0.0384**
Щука	6	29	9	4	4	52	34.2	8.207	0.0837
Язь	**7**	**35**	**15**	**13**	**13**	**83**	**54.6**	**12.26**	**0.0156**

Примечание: виды, для которых обнаружены достоверно значимые отличия во встречаемости между участками водохранилища, выделены жирным шрифтом.

Проанализируем предварительно "перекос" частот встречаемости отдельных видов на различных участках с применением критерия χ^2. Оценку *p*-значений выполним с использованием рандомизационного теста, как это мы делали ранее, поэтому наличие частот вида $n_j = 0$ или $n_j < 5$ можно считать допустимыми. Очевидно, что

пространственная изменчивость некоторых видов имеет высокую статистическую значимость (см. табл. 3.3), а другие виды являются независимыми или взаимозависимыми.

Но как выявить общие различия в видовой структуре сообществ как целостных совокупностей, образуемых видами? Здесь необходим многомерный анализ, обобщающий все переменные, каждая из которых соответствует тому или иному виду, составляющему ихтиоценоз. В терминах анализа сообществ это означает, что в единой процедуре должна быть рассмотрена вся совокупность видов, как структурный элемент экосистемы, обладающий эмерджентными надпопуляционными свойствами. Одним из возможных подходов может явиться многомерный дисперсионный анализ в пространстве численностей видов (или иных количественных характеристик обилия), но возможности параметрических процедур существенно ограничиваются проблемой исполнения исходных предпосылок о нормальном законе распределения данных.

Рассмотрим в этой связи другую задачу: имеются ли различия между отдельными участками водохранилища во всей совокупности видовой структуры рыбных сообществ? Решение ее возможно, например, (а) с использованием точного метода Фишера (Fisher's Exact Test) и (б) непараметрического анализа комбинаций одномерных статистик.

Точный метод Фишера осуществляет перебор всех возможных комбинаций таблиц сопряженности с одними и теми же маргинальными частотами, что и эмпирическая матрица (т.е. всех вариантов нуль-модели **FF**). При этом подсчитывается, какой процент таких таблиц содержит частоты, более резко отличающиеся от нулевого случая, чем исходная. Фишер показал, что вероятность получения любого набора частот n_{ij} таблицы $r \times c$ задается гипергеометрическим распределением:

$$p = \frac{(R_1!R_2!...R_r!)(C_1!C_2!...C_c!)}{N!\prod_{i,j} n_{ij}!}$$, где R_i и C_j - суммы по строкам и столбцам соответственно.

Разумеется, при больших значениях частот и размерностях таблицы число возможных комбинаций нуль-модели достигает астрономических величин. К счастью, функция fisher.test() статистической среды R имеет возможность запустить процесс Монте-Карло и оценить *p*-значение для нулевой гипотезы по ограниченному подмножеству B случайных реплик. Можно отметить хорошую сходимость результатов рандомизационного теста: при увеличении параметра B от 1000 до 100000 полученное *p*-значение варьировало в пределах от 0.0002 до 0.0001, т.е. нулевую гипотезу об одинаковой предрасположенности рыбного сообщества к разным участкам водохранилища можно уверенно отклонить. Отметим, что здесь, как и во всех остальных случаях, мы использовали двусторонний тест: при рандомизации не доставляет никаких проблем всегда подсчитывать число случаев, когда абсолютное значение эмпирической величины критерия оказалось ближе к 0, чем нуль-модельные реплики.

Непараметрический анализ комбинаций одномерных тестов имеет ряд преимуществ многомерного подхода и находит широкий диапазон применения. Одна из версий комбинаторного анализа одномерных статистик в рамках пермутационной процедуры, основанная на идеях обобщающего теста (omnibus test – Good, 2005б, p. 170), была реализована для этого примера В.Н. Якимовым:

1. Осуществляется генерация большого числа (B = 1000) нуль-моделей FF размерностью $r \times c$ = 34×5 и для каждой i-й переменной (т.е. каждого вида рыб из 34) и каждой j-й нуль-модели рассчитывается значение тест-статистики G (в рассматриваемом случае – χ^2). Формируется матрица **G** размерностью $r \times (B + 1)$ = 34×1001, первый столбец которой составляют значения G_{obs}, полученные по экспериментальным данным.

2. Отдельно в каждой из строк выполняется процедура ранжирования $R_{ij} = R(G_{ij}) = \sum_{k=0}^{B+1} I[G_{ik} \le G_{ij}]$, т.е. из матрицы **G** формируется матрица **R** из рангов G, полученных при переборе всех нуль-моделей. Тем самым мы оцениваем, какое место займет эмпирическое значение G_{obs} в ряду тест-статистик, полученных при справедливости нулевой гипотезы.

3. Для каждой из $B + 1$ совокупностей рангов вычисляется комбинирующая функция Фишера (Fisher's omnibus statistic): $W_j = -\sum_{i=1}^{r} \ln\left[\frac{B + 0.5 - R_{ij}}{B+1}\right]$.

4. Итоговое *p*-значение рассчитывается аналогично одномерному пермутационному тесту $p = \frac{0.5 + \sum_{j=1}^{B} I[W_j \geq W_{obs}]}{B+1}$, где W_{obs} – статистика для экспериментальных данных.

Многомерная версия перестановочного теста для всего видового списка рыб дает наблюдаемое значение комбинирующей функции Фишера W_{obs} = 62.2. Распределение этой статистики для всех 1000 нуль-моделей (т.е. при условии отсутствия каких-либо отличий между участками водохранилища) имеет 95% доверительный интервал от 23.9 до 42. Вероятность получить такое эмпирическое значение при справедливости нулевой гипотезы крайне мало и на основе полученного распределения составляет p = 0.001. Таким образом, можно говорить о высокой степени достоверности различий в видовой структуре ихтиоценозов между пятью участками водохранилища, зафиксированных на основе данных о встречаемости видов.

К разделу 3.2:

```
#  Рандомизация и таблицы сопряженности
source("print_rezult.r")  # Загрузка функций вывода результатов
# Определение таблицы с исходными данными
Gast <- matrix(c(
270,  83,   15,   83,  2,    36,  1,    24,  4,    1,  3,  1,    0,  1,
448,  126,  69,  18,  15,  5,    22,  34,  48,  2,  0,  17,  0,  0,
639,  369,  58,  79,  26,  36,  23,  78,  37,  0,  0,  0,    4,  0,
173,  342,  6,    5,    6,    0,     0,   1,    1,    0,  0,  0,    3,  1)
, nrow=14, dimnames = list(Species =
c("12345","1(23)45","10345","12045","12305","10045","12005","10005","10305","02045","00005","
00000","1(23)(45)","123(45)"),
Sites = c("Симферопольская","Южная","Степная","Керченская")))
# Компоненты анализа, постоянные для всех критериев и итераций
rowTotals <- rowSums(Gast)  #  сумма частот по строкам
colTotals <- colSums(Gast)  #  сумма частот по строкам
nOfCases <- sum(rowTotals)  #  объем выборки
Nrand = 1000
#  ------------- Анализ мер независимости и связи
#   Тест  χ² Пирсона с использованием функции из пакета stat
chisq.test(Gast)  #   С использованием асимптотического приближения
chisq.test(Gast, simulate.p.value = TRUE, B = Nrand) #   С использованием рандомизации
#   Собственная функция расчета статистики χ² (m - исходная матрица частот)
chisq <- function(m) {
  ex <- margin.table(m, 1) %o% margin.table(m, 2) / sum(m) ;   sum((m - ex)^2 / ex) }
#   Рандомизационный тест. Используем функцию r2dtable() из пакета stat
simchi <- sapply(r2dtable(Nrand, rowTotals, colTotals), chisq)
RandRes(chisq(Gast), simchi, Nrand)
#   Тест с использованием коэффициента  τ Валлиса-Гудмена-Краскела
#   Функция расчета τ столбцы|строки  (dat - исходная матрица частот)
tau.CR <- function(dat)  {
       uncond <- nOfCases^2;    for(i in 1:nrow(dat))
       uncond <- uncond-nOfCases*sum(dat[i,]^2/rowTotals[i]);
       cond <- nOfCases^2-sum(colTotals^2); return (1-(uncond/cond)) }
       simtau.CR <- sapply(r2dtable(Nrand, rowTotals, colTotals), tau.CR)
       RandRes(tau.CR(Gast), simtau.CR, Nrand)
#   Функция расчета τ строки|столбцы
tau.RC <- function(dat)  {
       uncond <- nOfCases^2;   for(j in 1:ncol(dat))
       uncond <- uncond-nOfCases*sum(dat[,j]^2/colTotals[j]);
       cond <- nOfCases^2-sum(rowTotals^2);
       return (1-(uncond/cond)) }
```

```
simtau.RC <- sapply(r2dtable(Nrand, rowTotals, colTotals), tau.RC)
RandRes(tau.RC(Gast), simtau.RC, Nrand)
#   Тест с использованием коэффициента  λ  Гудмена-Краскела
#   Функция расчета τ столбцы|строки  (dat - исходная матрица частот)
# Аргумент Ind определяет независимую переменную: по строкам Ind =2 или по столбцам Ind =1
lambda <- function(dat, Ind = 1) {
   L <- (sum(apply(dat, Ind, max)) - max(rowSums(dat))) / (sum(dat)-max(rowSums(dat)))
   return (as.numeric(L))}
empLambda <- lambda(Gast)
simLambda <- sapply(r2dtable(Nrand, rowTotals, colTotals), lambda)
RandRes(lambda(Gast), simLambda, Nrand)
#   Тест с использованием рангового критерия  τc Кендалла
#  число возрастающих пар  (t - исходная матрица частот)
P = function(t) {
  r_ndx = row(t) ;   c_ndx = col(t)
  sum(t * mapply(function(r, c){sum(t[(r_ndx > r) & (c_ndx > c)])},  r = r_ndx, c = c_ndx))}
#  число убывающих пар
Q = function(t) {
       r_ndx = row(t)  ;   c_ndx = col(t)
       sum(t * mapply( function(r, c){sum(t[(r_ndx > r) & (c_ndx < c)])},
       r = r_ndx, c = c_ndx) )}
kendall_tau_c = function(t){
    t = as.matrix(t) ;  m = min(dim(t)) ;  n = sum(t)
    return(ks_tauc = (m*2 * (P(t)-Q(t))) / ((n^2)*(m-1))) }
simtau <- sapply(r2dtable(Nrand, rowTotals, colTotals), kendall_tau_c)
RandRes(kendall_tau_c(Gast), simtau, Nrand)
#  ------------- Энтропийный дисперсионный анализ
#   Предварительные определения
DFA <- ncol(Gast)-1 ; library(vegan)
#  Функция расчета индекса Шеннона (логарифм с основанием 2)
diversity2 <- function(x)  diversity(x,base=2)
N0 <-(nOfCases  - sum(colTotals*colTotals)/nOfCases)/DFA   #  средние частоты по столбцам
T <- nOfCases*diversity2(rowTotals) #  Индекс Шеннона для всех морф
# ------- функция выполнения ЭДА-анализа
EDA <- function (data, invar=1) {
# При  invar=1 - возвращается Eta, при  invar=2 - таблица дисперсионного анализа
# Рассчитываем индексы Шеннона по местообитаниям
   colH <- apply(data, 2, diversity2) ; R <- sum(colTotals*colH)
   F <- ((T-R)/DFA)/(R/(nOfCases-DFA-1)) ; Eta <- (F-1)/(F-1+N0)
       if (invar == 1)  return (Eta)
       else if (invar == 2)  {
# Помещаем результаты в таблицу дисперсионного анализа
  FackTable<- matrix(replicate(3*5, NA), nrow=3)
  rownames(FackTable) <- c("Между групп", "Внутри групп","Общие")
  colnames(FackTable) <- c("Сумма квадратов", "Степеней свободы", "Средние квадраты",
                                        "F-отношение","Эта")
  FackTable[3,1]= T; FackTable[2,1] = R ; FackTable[1,1] = T-R
  FackTable[1,2]= DFA; FackTable[2,2] = nOfCases - DFA -1 ; FackTable[3,2] = nOfCases - 1
  FackTable[,3]= FackTable[,1]/FackTable[,2]; FackTable[1,4]=F ; FackTable[1,5]=Eta }
       return (FackTable) }
EDA(Gast, 2)
# --- Выполнение рандомизационного теста
simEta <- sapply(r2dtable(Nrand, rowTotals, colTotals), EDA)
RandRes(EDA(Gast), simEta, Nrand)
# ---------------- Многомерная версия пермутационного теста (омнибусный тест)
#   Перенос частот встречаемости рыб (см. табл. 3.3) через буфер обмена в таблицу R
#   Fish <-read.delim("clipboard", row.names=1)
#   Или загрузить их из сохраненного двоичного файла
load("fish.RData")
#   Точный тест Фишера для таблицы 5 × 34 с вычислением p-значения методом Монте-Карло
fisher.test(Fish, alternative="two.sided", simulate.p.value=TRUE, B=100000)
#   Определение базовых переменных
rowTotals <- rowSums(Fish)  #  сумма частот по строкам
```

```
colTotals <- colSums(Fish)  #  сумма частот по строкам
nOfCases <- sum(rowTotals)  #  объем выборки
Ncol = ncol(Fish) ; Nrow = nrow(Fish)
total <-    c(23, 61, 33, 18, 17)     # общее кол-во неводных съемок
Nrand=1000                                  # задаваемое число итераций
#   Функция расчета частных значений Хи-квадрат
chisq_V <- function(X) {
    m <-t(rbind(X,total-X)) ; ex <- margin.table(m, 1) %o% margin.table(m, 2) / sum(m)
    sum((m - ex)^2 / ex)  }
#    функция для комбинирующей (омнибусной) формулы Фишера
B <- function(Y) { -sum(log((Nrand + 1.5 - Y)/(Nrand + 2))) }
#    Создание матрицы частных значений Хи-квадрат
Vran <- Rran <-matrix(rep(NA, Nrow*(Nrand+1)), nrow=Nrow)
for (i in 1:Nrow) Vran[i,1] = chisq_V(Fish[i,])
for (j in 1:Nrand) {
       #   Генерация экземпляра нуль-модели FF
             Rm<-r2dtable(1, rowTotals, colTotals)[[1]]
             for (i in 1:Nrow) Vran[i,j+1] = chisq_V(Rm[i,]) }
#    Ранжирование частных критериев
for (i in 1:Nrow) Rran[i,] <- rank(Vran[i,])
#    Расчет критерия для омнибусного теста и вывод результатов рандомизации
Bf <- apply(Rran, 2, B)  ;  RandRes(Bf[1], Bf[2:(Nrand+1)] , Nrand)
```

3.3. Статистическая значимость регрессии двух переменных

Модели регрессии используются в практике эколого-биологических исследований для решения широкого круга следующих задач:

1. Компактное *описание*, позволяющее представить наиболее объяснимые с предметной точки функциональные отношения между переменными в рамках формальной модели и найти оценку ее параметров, основанную на эмпирических данных.

2. *Обобщение* результатов выборочных наблюдений и распространение статистических выводов на всю генеральную совокупность (в форме, например, доверительных интервалов модели).

3. *Прогноз* (или предсказание), заключающийся в вычислении значений переменной отклика с использованием уравнения регрессии.

Пусть мы имеем совместное (не обязательно нормальное) распределение двух случайных величин (*X, Y*). Предположим, что анализируемые данные получены из эксперимента, где случайная величина *X* считается *независимой* (т.е. не испытывающей влияние посторонних факторов), а *Y* – переменная *отклика* (response), которая может зависеть как от *X*, так и от множества других неучтенных факторов. Эксперимент спланирован таким образом, что способ распределения реализаций *X* по объектам наблюдений не вносит дополнительной систематической ошибки при оценке *Y*.

Под *регрессией Y* на *X* понимается зависимость, задающая траекторию движения точки (y_x, *x*), которая для каждого текущего значения *x* определяется условным математическим ожиданием $y_x = \mathrm{E}(Y|X = x)$. Эта траектория моделируется с использованием некоторой *функции регрессии* $f(x, \theta)$ с постоянными параметрами θ, которая позволяет оценить средние значения $\hat{y}$ реализаций зависимой переменной *Y* для каждого фиксированного значения *x* независимой переменной (предиктора) *X*. Поскольку для точного описания $f(x, \theta)$ необходимо знать закон условного распределения $\mathrm{E}(Y|X = x)$, то на практике ограничиваются ее наиболее подходящей аппроксимацией $\hat{f}_a(x,\theta)$, доставляющей минимум ошибки ε восстановления значений $Y(x)$.

Если характер наблюдаемых данных не противоречит этому, то первым естественным приближением функции регрессии является линейная зависимость и тогда

функциональное соотношение между наблюдаемыми значениями y и x называется *моделью линейной регрессии*:

$$y = \mathrm{E}(Y|X = x) + \varepsilon = \beta_0 + \beta x + \varepsilon_y, \qquad (3.1)$$

где β_0 и β – параметры или коэффициенты функции регрессии; ε_y – случайные остатки, в отношении которых делается предположение, что они статистически независимы (т.е. для каждой пары реализаций i, j $\mathrm{E}\varepsilon_i\varepsilon_j = 0$, $\forall i \neq j$) и одинаково распределены. Для так называемой "классической" линейной модели вводятся дополнительные и весьма жесткие предположения о стохастической структуре модели:

° ошибки ε_i нормально распределены с нулевым средним, $\mathrm{E}\varepsilon_i = 0$;

° дисперсия остатков ε_i одинакова для всех x-ов на всем диапазоне наблюдаемых данных (homoscedasticity), $\mathrm{E}\varepsilon_i^2 = \sigma^2$, $\varepsilon_i \sim \mathrm{N}(0, \sigma_y^2)$.

Пусть необходимо выполнить линейный регрессионный анализ с использованием выборки из пар наблюдений (x_1, y_1), (x_2, y_2), ..., (x_n, y_n). Первый и основной этап заключается в построении эмпирического уравнения $\hat{y}_i = b_0 + bx_i$, $i = 1, 2, \ldots, n$; т.е. получении таких оценок b_0 и b неизвестных параметров β_0 и β, которые наилучшим образом "подогнаны" (fitting) под выборочные экспериментальные данные. Проверке статистических гипотез относительно значимости оценок коэффициентов b_0 и b предшествуют пункты стандартного протокола валидации регрессии, который заключается в анализе нарушений предположений классической модели, таких как:

° стохастичность (недетерминированность) значений предиктора, которая может проявиться в коррелированности регрессора и остатков $\mathrm{E}\{\varepsilon|x\} \neq 0$ и, как следствие, привести к смещенным оценкам коэффициентов модели;

° "странности" в характере распределения остатков и их гетероскедастичности (т.е. неоднородности дисперсии σ_ε^2 в зависимости от x), что приводит к неверной оценке уровней значимости коэффициентов;

° автокоррелированность остатков при анализе временных рядов;

° наличие "выбросов" (значений, аномально отклоняющихся от общей тенденции);

° нелинейный характер функции регрессии $\mathrm{E}\{y|x\}$.

Корректная процедура регрессионного анализа заключается в диагностике перечисленных нарушений и их исправлении, т.е. подборе такой формы регрессии, которая бы наилучшим образом компенсировала выявленные отклонения. Общий оптимизационный подход реализуется с точки зрения минимизации функции потерь $\rho(\hat{\varepsilon})$, комбинирующей различным способом невязки между прогнозируемыми и фактическими значениями: $\sum_{i=1}^{n} \rho[(y_i - x_i b)] = \sum_{i=1}^{n} \rho(\hat{\varepsilon}) \to \min$.

Обратим внимание на наиболее важные частные случаи функции потерь $\rho(e)$ (Айвазян, Мхитарян, 1998):

° $\rho(\hat{\varepsilon}) = \hat{\varepsilon}^2$ приводит к среднеквадратичной регрессии, а метод, реализующий минимизацию суммы квадратов остатков модели $SS_e = \sum_{i=1}^{n} (y_i - \hat{y})^2$, принято называть *методом наименьших квадратов* (МНК или LSQ – Least Squares regression);

° $\rho(\hat{\varepsilon}) = |\hat{\varepsilon}|$, т.е. минимизируется сумма абсолютных разностей $|y_i - \hat{y}|$ *методом наименьших модулей* (Least Absolute Deviations – LAD), а соответствующую регрессию называют медианной.

Если выборочные ошибки модели регрессии $\hat{\varepsilon} = \hat{\varepsilon}_1, \hat{\varepsilon}_2, \ldots, \hat{\varepsilon}_n$ распределены нормально, то оценивание параметров методом наименьших квадратов LSQ является несмещенным и наиболее эффективным. Этот метод имеет также несомненное преимущество в том, что обеспечивается однозначное решение, поскольку для центрированных x_c и y_c оценка коэффициента β сводится к нахождению отношения ковариации к выборочной дисперсии независимой переменной (см. раздел 3.1):

$$b = \mathrm{E}\{x_c\, y_c\}/\mathrm{E}\{X^2\} = Cov(x_c\, y_c) / Sx_c; \quad y_c = x_c\, Cov(x_c\, y_c) / Sx_c + \hat{\varepsilon}.$$

В противном случае наличие асимметрии и "длинных хвостов" $\hat{\varepsilon}$ обычно вызывают обоснованное беспокойство. Возможный и далеко не самый лучший метод борьбы с этим явлением – удалить наибольшие остатки как выбросы и повторить расчеты. Альтернативой является использование различных вариантов робастной (robust) и устойчивой (resistant) регрессии. Рассмотрим два из них.

Регрессия Хубера (Huber) пытается найти компромисс между двумя крайними случаями LSQ и LAD:

- небольшие остатки возводятся в квадрат $\rho(\hat{\varepsilon}) = \hat{\varepsilon}^2$ при $|x| < c$;
- для больших отклонений при $|x| > c$ используются простые разности $\rho(\hat{\varepsilon}) = 2c|\hat{\varepsilon}| - c^2$.

Пороговое значение c (tuning constant) находится с использованием различных эвристик (например, $c = 1.345S_e$, где S_e – стандартное отклонение для остатков) .

Другой известный подход основан на методе урезанных квадратов (LTS - Least Trimmed Squares), в котором минимизируется $\sum_{i=1}^{q} \hat{\varepsilon}^2_{(i)}$, где $q < n$, а (i) – комбинации наилучших индексов x, которые подбираются генетическим алгоритмом (Jung, 2007).

Другие полезные версии робастных регрессий будут нами рассматриваться далее по мере изложения. Таким образом, поиск модели регрессии, которая наилучшим образом описывает двумерное облако экспериментальных точек, является фундаментальной проблемой анализа данных, не имеющей пока конечного решения.

Рассмотрим предварительно геометрическую интерпретацию двух основных типов моделей регрессии (Sokal, Rohlf, 1981; Legendre, Legendre, 1998).

Модель I типа применяется в условиях, когда объясняющая переменная X является управляемой (т.е. известной априори) или случайная составляющая ее вариации существенно меньше ошибки Y. Кроме гипотезы, что переменные в эксперименте линейно связаны, для модели I существенны все предположения классической модели. Оценки параметров регрессии β и β_0 легко найти по формулам наименьших квадратов LSQ:

- для коэффициента b это тангенс угла наклона линии регрессии в координатах xy

$$b = \{\sum y_i x_i - \bar{x}\sum y_i\} / \{\sum x_i^2 - (\sum x_i)^2 / n\}; \qquad (3.2)$$

- для коэффициента b_0 – постоянный отрезок, отсекаемый этой линией на оси ординат

$$b_0 = \{\sum y_i - b\sum x_i\} / n.$$

Концептуально модель регрессии I-го типа ориентирована на *оптимальный прогноз* значений отклика Y. Геометрически при этом ищется минимум суммы квадратов разностей ординат точек, т.е. отсчет ведется строго по вертикали – см. рис. 3.3а. Можно также показать, что найденная прямая $Y = f(X)$ проходит через "центр тяжести" графика, образованный средними значениями переменных ($\bar{x}, \bar{y}$).

Для модели I-го типа гипотеза корреляции (то есть взаимозависимости) бессмысленна, т.к. X по определению является независимой переменной. Однако на практике через центральную точку ($\bar{x}, \bar{y}$) проходит и линия обратной регрессии $X = f(Y)$, а коэффициент корреляции $R(X, Y)$, который не равен 1, является геометрическим средним из оценок параметров регрессии каждой из переменной на другую: $c_{(X \cdot Y)} = R^2 / b_{(Y \cdot Y)}$ (в обозначениях рис. 3.3а) или $R = tg(45^o - \theta/2)$. Величина угла θ между двумя линиями регрессии может быть использована для оценки нелинейных нарушений функции $\mathrm{E}\{y|x\}$. Подробное обсуждение аспектов этих непростых парадоксов см. в книге (Legendre, Legendre, 1998, p. 503).

Модели *регрессии II типа* используются, когда требуется оценить параметры уравнения, описывающего функциональные отношения между парой двух случайных переменных (т.е. уже X не является независимой или управляемой). Если случайные вариации X и Y равновелики и пропорциональны, то на оценку параметра β угла наклона прямой, найденную обычным МНК (LSQ), оказывает влияние наличие ошибки измерения

объясняющей переменной. Тогда можно предположить, что модели II более корректным образом будут оценивать *степень отношения* между двумя исходными переменными (Sokal, Rohlf, 1981).

а) Модель наименьших квадратов (LSQ) б) Модель главных осей регрессии (MA)

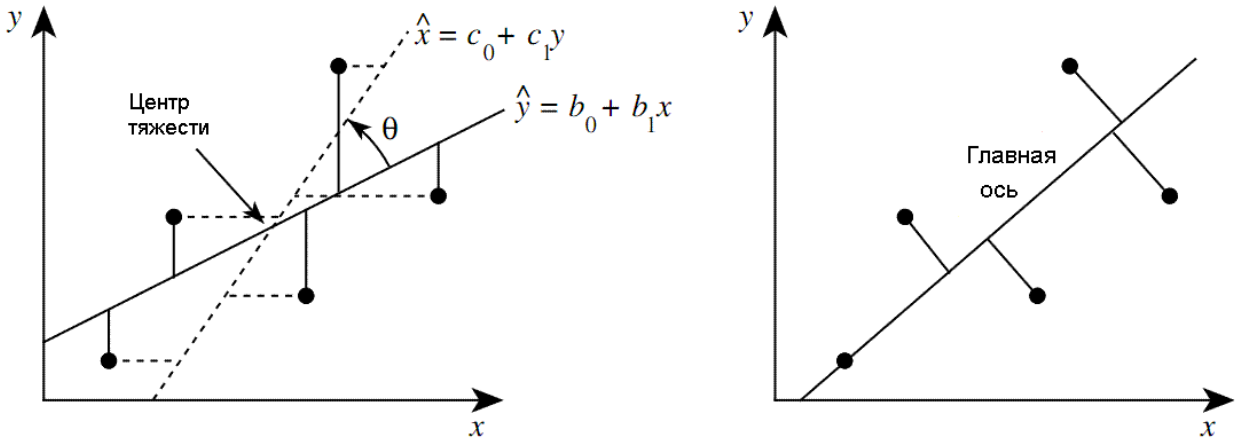


Рис. 3.3. Модели регрессии I типа (а) двух случайных переменных x и y : для $y = f(x)$ суммируются и минимизируются квадраты отклонений по вертикали (сплошные линии); для модели $x = f(y)$ – квадраты отклонений по горизонтали (пунктир). Для модели регрессии II типа (б) суммируются и минимизируются квадраты евклидовых дистанций до линии регрессии (т.е. главной оси)

При аппроксимации данных моделями II имеет значение направление, в котором измеряются отклонения от линии регрессии:

- в методе главных осей (MA – Major Axis regression) расстояния от эмпирических точек отсчитываются в перпендикулярном направлении к аппроксимирующей прямой (см. рис. 3.3б);
- в стандартном или редуцированном методе главных осей (RMA – Reduced Major Axis) разности находятся вдоль направления, которое является отражением линии регрессии относительно оси Y;
- ранжированный метод главных осей осуществляет предварительную деформацию осей переменных для достижения однородных рангов.

Коэффициенты этих уравнений, основанные на геометрических представлениях, рассчитываются по простым формулам, например, для модели RMA:

$$b_{RMA} = \sqrt{\frac{\sum y_i^2 - (\sum y_i)^2 / n}{\sum x_i^2 - (\sum x_i)^2 / n}}, \; b_{0\,RMA} = \bar{y} - b_{RMA}\bar{x}.$$

Для представленного в разделе 3.1 примера [П2] уравнения регрессии Y (доля донных организмов Diamesinae+Orthocladiinae d_{DO}) на X (температура t), рассчитанные по представленным формулам, будут иметь вид (см. рис. 3.4):

$Y = 135 - 4.98\,X$ (модель I, метод LSQ, $\theta = -78.6$);
$Y = 174 - 7.34\,X$ (модель II, метод RMA, $\theta = -83.8$);
$Y = 229 - 10.7\,X$ (модель II, метод MA, $\theta = -84.7$).

Подробные рекомендации, в каких условиях какая из моделей предпочтительнее, основанные на соотношении оценок дисперсий σ_X, σ_Y, σ_ε, и σ_δ, обосновываются П. и Л. Лежандрами (1998). Мы только обратим внимание читателя, что при использовании моделей II типа угол θ между линиями прямой и обратной регрессии увеличивается.

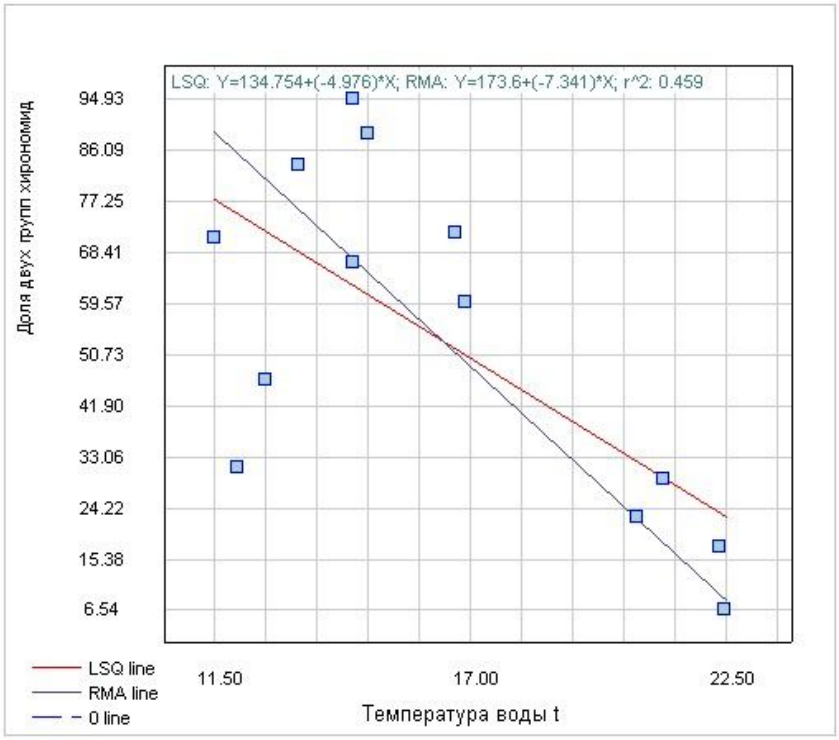


Рис. 3.4. Линейная регрессия доли Diamesinae+Orthocladiinae на температуру воды

Для примера на рис. 3.4 уравнения робастной регрессии будут иметь вид:

$Y = 147.3 - 5.62\,X$ (регрессия Хубера);

$Y = 146.9 - 5.61\,X$ (метод LTS).

Различия в коэффициентах b найденных уравнений выглядят вполне умеренными, но хотелось бы получить точный ответ на вопрос, какое из уравнений предпочтительней.

Статистический анализ значимости полученных моделей регрессии параметрическими методами осуществляется с использованием нескольких критериев. Отметим, что все соотношения, приведенные ниже, справедливы лишь при предположении о гауссовом характере распределения остатков **ε**. Чтобы проверить, равна ли оценка коэффициента регрессии b истинному значению β, обычно выполняют анализ на равенство нулю статистической величины $(b - \beta)/s_b$, сравнивая ее с t-распределением, имеющим $n - 2$ степеней свободы. По сути, это идентично выяснению H_0: $\beta = 0$, т.е. превышает ли абсолютная величина тангенса угла наклона b достаточно малое гипотетическое значение. Здесь s_b – стандартная ошибка b, $s_b = s_e \Big/ \sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}$; s_e – стандартное отклонение для остатков, $s_e = \sqrt{\frac{1}{n-k}\sum_{i=1}^{n}\left(y_i - a - bx_i\right)^2}$. Для нашего примера модели LSQ: $s_b = 1.627$, $t_{obs} = -3.05$, $p = 0.011$.

Для оценки статистической значимости уравнения регрессии в целом вычисляется дисперсионное отношение Фишера (Дрейпер, Смит, 1986), которое оценивает, насколько сумма квадратов отклонений SS_X, объясненная моделью, превышает остаточную сумму квадратов SS_e с учетом их степеней свободы:

$$SS_X = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \;;\; SS_e = \sum_{i=1}^{n}\left(y_i - \hat{y}\right)^2 \;;\; F = \frac{SS_X/(k-1)}{SS_e/(n-k)},$$

где $k = 2$ – число степеней свободы при простой линейной регрессии, $\hat{y}_i$ - расчетные значения. Если остатки $\varepsilon_1,\ldots,\varepsilon_n$ распределены нормально $N(0,\sigma^2)$, то F-статистика при

справедливости гипотезы H_0 имеет стандартное распределение Фишера с (k -1) и (n - k) степенями свободы. В нашем примере: F_{obs} = 9.35, p = 0.011, т.е. для простой регрессии этот тест идентичен проверке гипотезы H_0: β = 0 по t-критерию.

При сравнительной оценке стандартных ошибок выборочных коэффициентов для различных вариантов регрессий (в том числе, RMA, Хубера, LTS) удобно воспользоваться таким непараметрическим инструментарием статистических выводов, как рандомизация и бутстреп. В общем случае возможны два следующих варианта использования ресамплинга для анализа модели регрессии:

1. Оценка параметров. Повторные псевдовыборки по алгоритму "случайного выбора с возвращением" формируются из номеров строк исходного набора данных. В частности, исходная модель регрессии строится по данным, расположенным в исходной последовательности индексов {1, 2, …, 12, 13}. На первой итерации бутстрепа эта последовательность может приобрести, например, вид {7, 6, 12, 10, 9, 1, 9, 10, 6, 7, 10, 13, 4}, т.е. некоторые индексы исчезают, а другие повторяются два или несколько раз. Статистические связи между x и y в этом подмножестве строк полностью сохраняются. Регрессионная модель, построенная по такой перевыборке, может несколько отличаться от первоначальной, причем степень этих отличий зависит от степени ее устойчивости к легкой модификации исходных данных. Такие шаги бутстрепирования могут быть выполнены достаточно большое число раз, чтобы сформировать статистическое распределение коэффициентов модели или любых критериев качества аппроксимации.

Для этого примера используем алгоритм бутстрепа (Manly, 2007) с использованием остатков ε_i регрессионной модели, построенной на эмпирических данных: $y_i = \hat{y}_i + \hat{\varepsilon}_i$. Будем осуществлять многократно (B = 1000) извлечение случайных перевыборок из вектора остатков, получать новый вектор **Y*** и на основе его рассчитывать коэффициент b* для бутстрепированной регрессионной модели. Графики функций ядерного сглаживания распределений β для регрессий, полученных методами LSQ, Хубера и LTS, представлены на рис. 3.5. Бóльшее удаление от 0 доверительных интервалов для робастных моделей свидетельствует об их устойчивости по сравнению с моделью МНК.

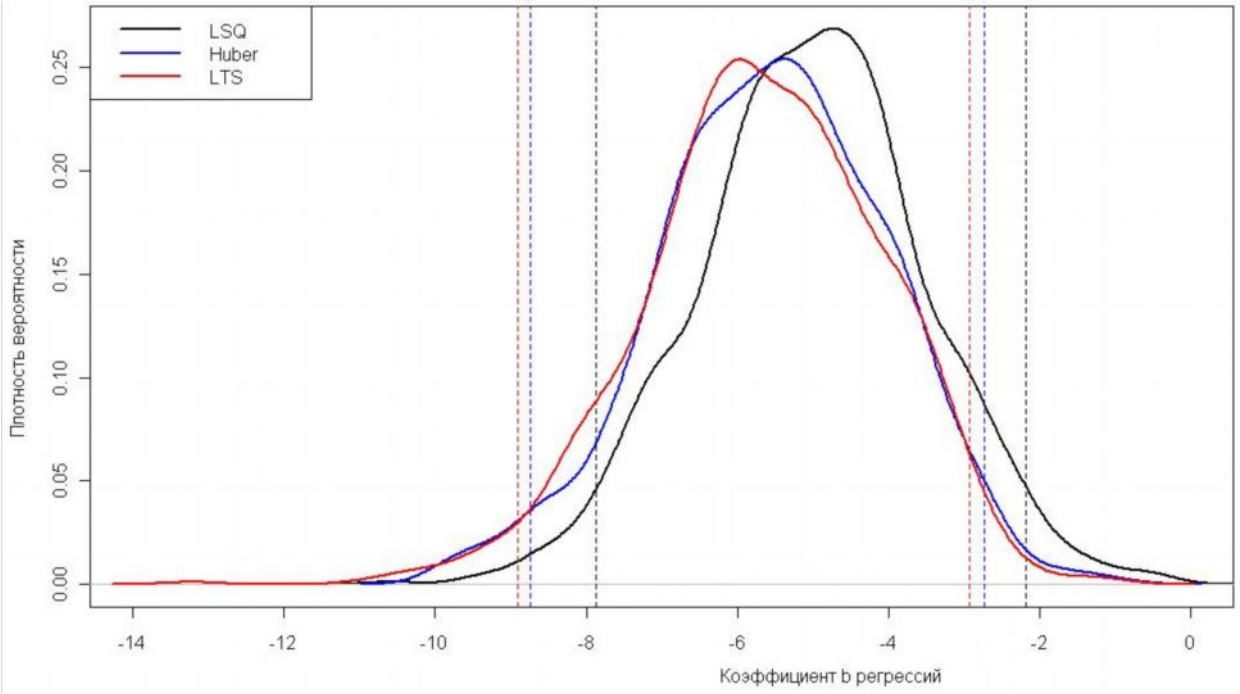


Рис. 3.5. Статистические распределения коэффициента *b* для различных моделей регрессии: по методу наименьших квадратов (LSQ), Хубера и методу урезанных квадратов (LTS)

Для моделей II типа нам нет необходимости проводить самостоятельных расчетов, поскольку использование бутстрепа заложено в самой функции lmodel2(…) пакета MASS

и 95% доверительные интервалы β имеют граничные значения -4.94 ÷ -27.23 для RMA и -6.21 ÷ -37.93 для MA.

2. Рандомизационный тест статистической значимости компонентов регрессионной модели не отличается от его версий, описанных выше (например, в разделах 2.4, 2.6, 3.1). В техническом плане для его реализации достаточно многократно (B = 1000) перемешивать по алгоритму "случайной перестановки" значения отклика Y (или предиктора X), каждый раз связывая пары x_i и y_i наугад. В результате этого статистические связи между переменной полностью разрушаются, т.е. можно получить распределение значений тестируемой статистики (b_{ran}, t_{ran} или F_{ran}) при справедливости нулевой гипотезы. Для рассматриваемого примера p-значения, полученные рандомизацией, незначительно отличаются от рассчитанных обычными параметрическими методами:

$$p(|b_{ran}| \le |b_{obs}|) = 0.0152;\ p(|F_{ran}| \le |F_{obs}|) = 0.0119.$$

К разделу 3.3:

```
#    Модели регрессии 1-го и 2-го типа
#   Используем те же данные, что и для раздела 3.1
NDO <- c(88.9,94.9,70.8,46.4,31.0,66.5,83.6,71.9,59.6,22.5,29.2,6.5,17.5)
T <- c(14.8,14.5,11.5,12.6,12,14.5,13.3,16.7,16.9,20.6,21.2,22.5,22.4)
source("print_rezult.r")  # Загрузка функций вывода результатов
#    Использование простой линейной модели
Ex1 <-lm(NDO ~ T) ; Ex0 <-lm(NDO ~ 1) ; summary(Ex1); anova(Ex1, Ex0)
#  Получение F-статистики и углового коэффициента для эмпирических данных
F_obs <- anova(Ex1, Ex0)$F[2] ; b_obs <- Ex1$coefficients[2]
Nperm = 1000 ; F_rand <- b_rand <- as.numeric(rep(NA, Nperm))
for (i in 1:Nperm) {
       #   Рандомизация Y и построение модели на рандомизированных данных
       NDO_rerm <- sample(NDO,length(NDO)) ;  Ex_rand <- lm(NDO_rerm ~ T)
       F_rand[i] <- anova(Ex_rand)$F[1] ; b_rand[i] <- Ex_rand$coefficients[2] }
RandRes(F_obs, F_rand, Nperm) ;  RandRes(b_obs, b_rand, Nperm)  #   Вывод результатов
#  Сравнение различных типов моделей
library(lmodel2)     #   Подключаем пакет, вычисляющий различные модели 2-го типа регрессии
#   Для RMA-модели используем нормировку от значащего нуля. Задаем 999 пермутаций
Ex2.res <- lmodel2(NDO ~ T, range.y= "relative",range.x="relative",nperm=999) ; Ex2.res
#    Рисуем все 4 модели на одном графике
plot(Ex2.res, method="OLS", conf = FALSE, centroid=TRUE, main="",
                     xlab = "Температура, град", ylab = "Доля диамезин", col=1)
lines(Ex2.res, "SMA", col=2, conf = FALSE) ; lines(Ex2.res, "RMA", col=3, conf = FALSE)
lines(Ex2.res, "MA", col=4, conf = FALSE)   ;
legend("topright", c("OLS", "SMA", "RMA", "MA"), col=1:4, lty=1)
#    Рисуем все 4 графика для каждой модели модели, но с доверительными интервалами
op <- par(mfrow = c(2,2))  ; plot(Ex2.res, "OLS") ; plot(Ex2.res, "MA")
plot(Ex2.res, "SMA") ; plot(Ex2.res, "RMA") ; par(op)
library(MASS) ; summary(rlm(NDO ~ T))  # Робастная регрессия Хубера
library(robustbase) ; lmrob(NDO ~ T)  # Метод урезанных квадратов LTS
#  Бутстрепирование коэффициента b с использованием остатков
#  Имя функции, реализующей построение модели, передается как параметр
boot_dist <- function(Nboot,x,y, function_name) { FUN <- match.fun(function_name);
  g <- FUN(y ~ x) ;   n=length(x);  bcoef <- as.numeric(rep(NA, Nboot))
       for (i in 1:Nboot) {
         newy <- g$fit + g$res[sample.int(n,size = n, replace=TRUE)] ; brg <- FUN(newy ~ x )
         bcoef[i] <- brg$coef[2] }
       return (bcoef) }
bbLQ <- boot_dist(1000, T, NDO, "lm")
bbHu <- boot_dist(1000, T, NDO, "rlm")
bbLT <- boot_dist(1000, T, NDO, "lmrob")
#   Рисуем график ядерных функций на рис. 3.5
plot(density(bbLQ),xlab="Коэффициент b регрессий", xlim=c(-14,0), main="", lwd=2)
abline(v=quantile(bbLQ,c(0.025,0.975)), lty=2) ;
lines(density(bbHu),col="blue", lwd=2)
abline(v=quantile(bbHu,c(0.025,0.975)),col="blue",lty=2)
```

```
lines(density(bbLT),col="red", lwd=2) ;
abline(v=quantile(bbLT,c(0.025,0.975)),col="red",lty=2)
legend ("topleft", c("LSQ",  "Huber",  "LTS"),col = c("black","blue", "red"), lwd = 2)
```

3.4. Нелинейная регрессия и скользящий контроль

В разделе 3.3 общий вид модели регрессии от одной независимых переменной был представлен как $y = \mathrm{E}(Y|X = x) + \varepsilon = f(x, \theta) + \varepsilon$,
т.е. условное математическое ожидание отклика $\mathrm{E}(Y|\mathrm{X} = x)$ моделируется некоторой аппроксимирующей функцией $f(x, \theta)$ с постоянными параметрами θ, которая с ошибкой ε восстанавливает среднюю величину реализаций y случайной величины Y для произвольного значения x независимой переменной X. Этап структурной идентификации регрессионной модели сводится к выбору семейства функций $f(x, \theta)$, в рамках которых происходит дальнейшая оценка неизвестных параметров θ уравнения регрессии. Примерный состав аналитических выражений, обычно используемый для "подгонки" зависимостей, будет представлен далее в табл. 3.4.

Выбор алгебраической формы функции $f(x, \theta)$ далеко не всегда очевиден и осуществляется с учетом следующих обстоятельств:

- возможность непротиворечивой интерпретации регрессионной модели на основе теоретических представлений в конкретной предметной области;
- точность аппроксимации моделью имеющихся выборочных данных;
- устойчивость регрессионной модели при экстраполяции (т.е. при $x \to \infty$) и робастность по отношению к "выбросам";
- минимальная сложность ("экономность" модели), определяемая числом подбираемых параметров θ.

Привлечение априорной информации о механизме изучаемых процессов является часто решающим фактором при выборе формы зависимости. Например, всерьез принято считать, что зависимость "доза-эффект" в токсикологии описывается только пробит-функцией, а большинство аллометрических зависимостей между размерами двух структурных частей организма подчиняется степенному уравнению, предложенному Т. Гексли еще в 1932 г. Однако на практике вид функции $f(x, \theta)$ чаще всего нельзя однозначно задать, исходя только из эколого-биологических соображений.

Выбор конкретной модели (model selection) из достаточно небольшого числа моделей-претендентов является частным случаем более общей задачи определения структуры модели (structure selection). Естественным является желание выбрать модель, обеспечивающую минимальную ошибку ε, которая по мере увеличения сложности функции $f(x, \theta)$, как правило, монотонно убывает. Однако дополнительные члены, включаемые в модель, на определенном этапе уже не несут полезной информации или дублируют друг друга (что происходит, например, при бесконтрольном увеличении степени полинома). При этом средняя ошибка на независимых контрольных данных сначала уменьшается, затем проходит через точку минимума и далее только возрастает. Такие модели называют *переусложненными*.

Многие руководства по статистике предлагают ориентироваться на формальные критерии качества аппроксимации, такие как среднеквадратичная ошибка регрессии $s_e^2 = \frac{1}{(n-k)}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$ и коэффициент детерминации $R^2 = 1 - s_e^2 / s_y^2$, где k – число параметров модели, $\hat{y}_i$ – значения оцениваемого отклика, s_y^2 – оценка дисперсии зависимой переменной. Однако эти критерии слабо зависят от k, что часто приводит к переусложненным моделям, точно описывающим выборочные данные, но весьма неустойчивым при экстраполяции.

Современная технология выбора хорошо интерпретируемых моделей оптимальной сложности ориентируется на применение информационных критериев, основанных на функции максимального правдоподобия, и использовании методов перекрестной проверки.

Оценки метода наименьших квадратов являются оптимальными лишь при нормальном законе распределения остатков $\varepsilon_1,\ldots,\varepsilon_n$. Если нам известен только общий вид закона распределения вероятностей выборочных данных, чтобы найти набор наилучших оценок параметров $\boldsymbol{\theta}$, необходимо воспользоваться более универсальным способом – максимизировать логарифм функции правдоподобия (log-likelihood), описывающей "вероятность" наших данных для определенной модели:

$$LL = \ln\{L(\boldsymbol{\theta}|\boldsymbol{x}, \boldsymbol{y})\} \rightarrow \max.$$

Подробнее метод максимального правдоподобия рассмотрим ниже (раздел 3.6), а здесь остановимся на информационных критериях качества моделей, использующих *LL*.

Девианс G^2 для анализируемой модели M, основанной на наборе данных *y*, определен как $G^2 = -2\,(LL_M - LL_S)$ (Singer, Willett, 2003, 122 p.),

где LL_S – максимум логарифма функции правдоподобия для полной или "насыщенной" (saturated) модели S, которая содержит так много параметров $\boldsymbol{\theta}_S$, чтобы по возможности точно восстановить значения *y* выборочных данных. На практике вероятность точно воспроизвести данные для модели S обычно близка к 1, а логарифм максимума функции правдоподобия равен 0. Для тестируемых моделей M величина LL_M всегда отрицательна и тогда $LL_S > LL_M$ для любых M. Хотя статистическая величина девианса G^2 является неизвестной, ее оценка идентична выборочной дисперсии остатков $n\hat{\sigma}_e^2$, а распределение может быть аппроксимировано χ^2-распределением с *k* степенями свободы.

Информационный критерий IK добавляет к девиансу “штраф” за увеличение сложности модели и в общем виде определяется как $IK = G^2 + 2ck = -2\,(LL_M - 2ck)$,

где *k* – число параметров модели, *c* – масштабирующий множитель.

Для *информационного критерия Акаике* (Akaike information criterion) $c = 1$ и

$$AIC = G^2 + 2k = -2\ln\{L(\boldsymbol{\theta}|\mathbf{x},\mathbf{y})\} + 2k, = -n\ln(\hat{\sigma}_e^2) + 2k$$

При нормальном законе распределения остатков $AIC = -n\ln\left(\sum_{i=1}^{n}(\hat{y}_i - y_i)^2/n\right) + 2k$,

где *n* – количество наблюдений (Burnham, Anderson, 2002).

Для байесовского информационного критерия $c = \ln(n)$ и $BIC = G^2 + 2\ln(n)\,k$.

Можно упомянуть также скорректированный критерий AIC_c, где штраф за усложнение модели увеличен: $AIC_c = AIC + 2k(k+1)/(n-k-1)$.

Представленные выше показатели качества подгонки (goodness fit) относятся к *внутренним* критериям, т.е. они оценивают модель с использованием тех же данных, по которым эта модель была построена. Однако минимизация ошибок на обучающем множестве, которое не бывает ни идеальным, ни бесконечно большим, неизбежно приводит к моделям, смещенным относительно истинной функции процесса. Преодолеть это смещение можно только с использованием *внешних критериев.*

Само возникновение термина "внешний критерий" обязано аналогии с термином "внешнее дополнение" (Бир, 1963), что в свое время служило косвенным обоснованием предпочтительности их применения. Сейчас стало ясно, что, поскольку теоретически строгого определения обоих этих понятий не существует, представляется несколько неуместной ссылка на теорему Геделя о неполноте как аргумент целесообразности использования "принципа внешнего дополнения" (Розенберг и др., 1994). Однако если понимать под внешними критериями ситуацию, когда оценки параметров моделей θ и квадраты норм невязок вычисляются на несовпадающих множествах, то практически оказалось, что внешние критерии оказываются почти всегда лучше внутренних в смысле

различимости. Они также являются незаменимыми при выборе оптимального числа параметров моделей, как эффективное средство против их переусложнения.

Прямой путь к объективной оценке ошибки воспроизведения моделей – это использовать для проверки (validation) свежие, независимые данные из того же самого источника наблюдений. И если нам нужно выбрать одну модель из числа многих, мы выбираем ту, которая оказалась лучшей при испытании в независимых условиях. Поскольку получить новые порции данных чаще всего возможности не имеется, то будет логичным разделить доступные данные наугад на две части – обучающую и проверочную выборки, что является определенной гарантией того, что эти две выборки независимы и одинаково распределены. И, наконец, естественно сделать эту процедуру многократной.

Скользящий контроль или *перекрестная проверка* (cross-validation, CV) является общей процедурой эмпирического оценивания моделей, построенных по прецедентам. Для этого выполняется некоторое множество случайных разбиений исходной выборки на две подвыборки: обучающую и контрольную. Для каждого разбиения выполняется расчет параметров модели по обучающей подвыборке, после чего оценивается среднеквадратичная погрешность уравнения регрессии на объектах контрольной подвыборки. Если исходные выборки независимы, то средняя ошибка скользящего контроля даёт несмещённую оценку ошибки регрессии, и это выгодно отличает её от средней ошибки на обучающей выборке, которая при переусложнении модели может оказаться оптимистически заниженной.

Стандартной считается методика $t \times q$-кратного скользящего контроля ($t \times q$-fold cross-validation), когда выборка случайным образом t раз разбивается на q блоков равной длины, после чего каждый блок по очереди становится контрольной выборкой, а объединение всех остальных блоков – обучающей последовательностью. Частным случаем полного скользящего контроля является кросс-проверка с исключением одного объекта (leave-one-out CV), т.е. $t = q = n$. При этом строится n моделей регрессии по $(n - 1)$ выборочным значениям, а исключенная реализация каждый раз используется для расчета ошибки скользящего контроля.

Рассмотрим в качестве примера [П5] выбор уравнения регрессии для зависимости активности оксидазы *L*-аминокислот от протеолитической активности яда обыкновенной гадюки из различных популяций на территории Европейской части России (всего по 28 объединенным образцам). В качестве претендентов рассматривалось 4 модели регрессии, линейных по параметрам, для расчета коэффициентов которых использовалась функция gml(…) статистической среды R, а также 5 нелинейных моделей, оценка параметров которых осуществлялась с помощью функции nls(…). Кросс-проверка выполнялась на основе функции cv.gml(…) и других процедур, представленных скриптами в конце этого раздела. Критерии качества аппроксимации, учитываемые при селекции наилучшей модели, представлены в табл. 3.4.

По результатам табл. 3.4 можно отметить два обстоятельства. При переходе от полиномиальной модели с тремя статистически значимыми коэффициентами к аналогичному полиному 3-й степени (все четыре коэффициента которого уже незначимы) коэффициент детерминации и AIC оценили эти модели как совершенно равноценные. Однако оценка ошибки регрессии при скользящем контроле убедительно продемонстрировала низкую устойчивость полиномов высокой степени на данных, не участвовавших в построении моделей.

Интересно также поведение ошибки скользящего контроля при разных способах группировки. Например, если разделить данные на 7 блоков по 4 точки и выполнить q-кратный скользящий контроль, то для полинома 3-й степени ошибка возрастет до $s^2_{eCV} = 82$, что является дополнительным свидетельством недостаточной адекватности этой модели. Для остальных моделей ошибка скользящего контроля мало зависела от способа разбиения $t \times q$.

Таблица 3.4. Критерии качества аппроксимации данных по свойствам ядовитого секрета гадюки обыкновенной (s_e^2 – среднеквадратичная ошибка регрессии, R^2 – приведенный коэффициент детерминации, AIC – информационный критерий Акаике, s_{eCV}^2 – среднеквадратичная ошибка скользящего контроля)

Вид аппроксимирующей функции	s_e^2 (df=26)	R^2	AIC	s_{eCV}^2 (df=28)
Модели, линейные по параметрам				
Линейная $y = a + bx$	46.6	0.556	190.9	52.9
Гипербола $y = a + b/x$	64.4	0.387	200	191.5
Полином 2 степени $y = a + bx + cx^2$	35.3	0.664	184	44.6
Полином 3 степени $y = a + bx + cx^2 + dx^3$	35.1	0.665	184.8	61.9
Нелинейные модели				
Экспоненциальная $y = a\,e^{bx}$	59.5	0.456	197.8	73.8
Степенная $y = a\,x^b$	44.9	0.588	189.9	51.2
Логарифмическая $y = a + b\,x + c\,\ln(x)$	43.7	0.615	190	107.3
Экспоненциально-степенная $y = e^{ax}x^b$	38.7	0.649	185.7	42.8
Логистическая $y = a/(1+e^{b+cx})$	24.2	0.787	173.5	30.4

По всей совокупности критериев качества несомненное лидерство принадлежит следующей модели регрессии на основе логистической функции (см. рис. 3.6):

Вид модели	Коэффициенты	*t*-критерий	Pr(>|*t*|)
$y = \frac{25.48}{1-e^{8.96-0.86x}}$	a = 25.48	21.02	< 0.0001
	b = 8.96	2.634	0.014
	c = -0.86	-2.496	0.019

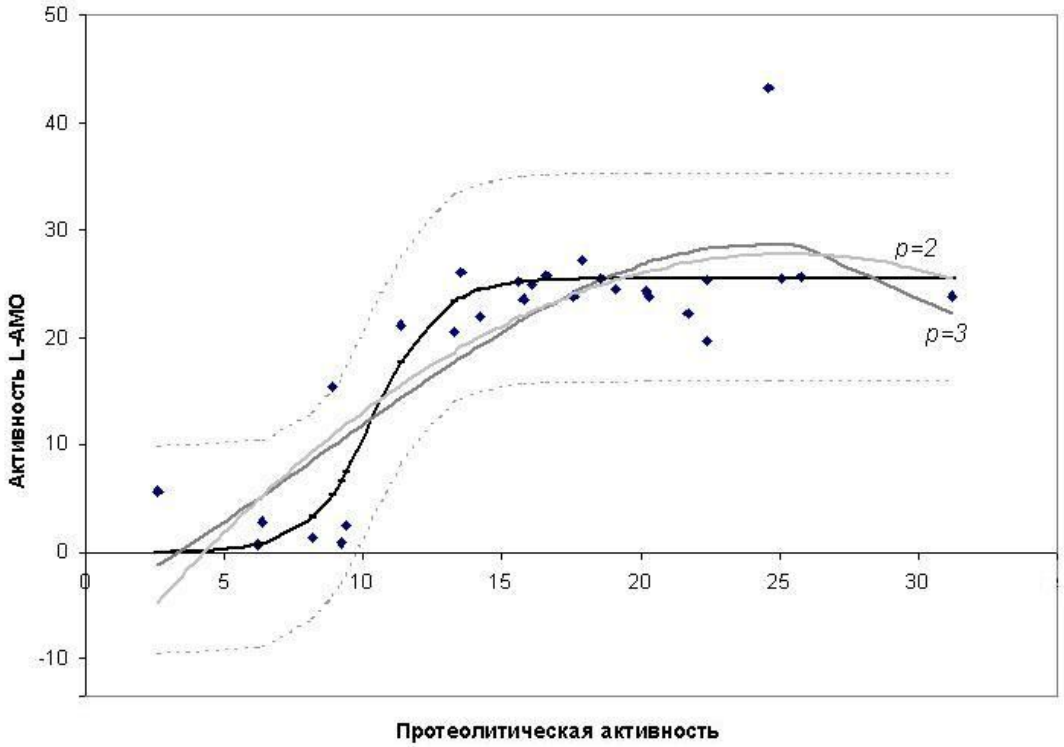


Рис. 3.6. Логистическая модель соотношения активности биохимических компонент ядовитого секрета гадюки (пунктиром показаны 95% доверительные интервалы, серым – для сравнения модели полиномов 2-й и 3-й степени). Расчеты сделаны программой AtteStat.

Представляет интерес построить на основе ресамплинга графики распределения критериев качества моделей, чтобы оценить их статистическую значимость и

доверительные интервалы. Как это сделать технически – описано в предыдущем разделе. К сожалению, провести бутстрепирование логистической модели с использованием функции nls(…) оказалось невозможным, т.к. ее построение требует каждый раз достаточно точной предварительной подстройки стартовых значений параметров, иначе алгоритм нелинейной оптимизации склонен к вырождению. Поэтому рассмотрим более простую модель полинома 2 степени, также имеющую неплохие показатели качества аппроксимации, но не имеющую вычислительных проблем при оценке коэффициентов.

В очередной раз обратим внимание на различие процедур бутстрепа и рандомизации. Гистограмма распределения критерия Акаике для полиномиальной модели, построенная в результате 400 итераций бутстрепа, представлена на рис. 3.7 (слева), а его 95% доверительный интервал находится в границах 149.4 ÷ 194.4.

Статистическое распределение AIC, полученное в ходе рандомизации (см. гистограмму на рис. 3.7 справа), соответствует нулевой модели процесса, т.е. если искомая зависимость будет отсутствовать. В этих условиях 95% доверительный интервал AIC-критерия имеет пределы 209.1 ÷ 216.7. Очевидно, поскольку доверительные интервалы AIC не пересекаются, это является свидетельством адекватности модели с точки зрения этого критерия. Поскольку у обеих гистограмм нет общих интервалов, приведенный анализ является одновременно тестом на статистическую значимость критерия Акаике: при отсутствии связи $y = f(x)$ мы не можем с вероятностью $p = 1/400$ получить AIC меньший, чем в эмпирическом случае.

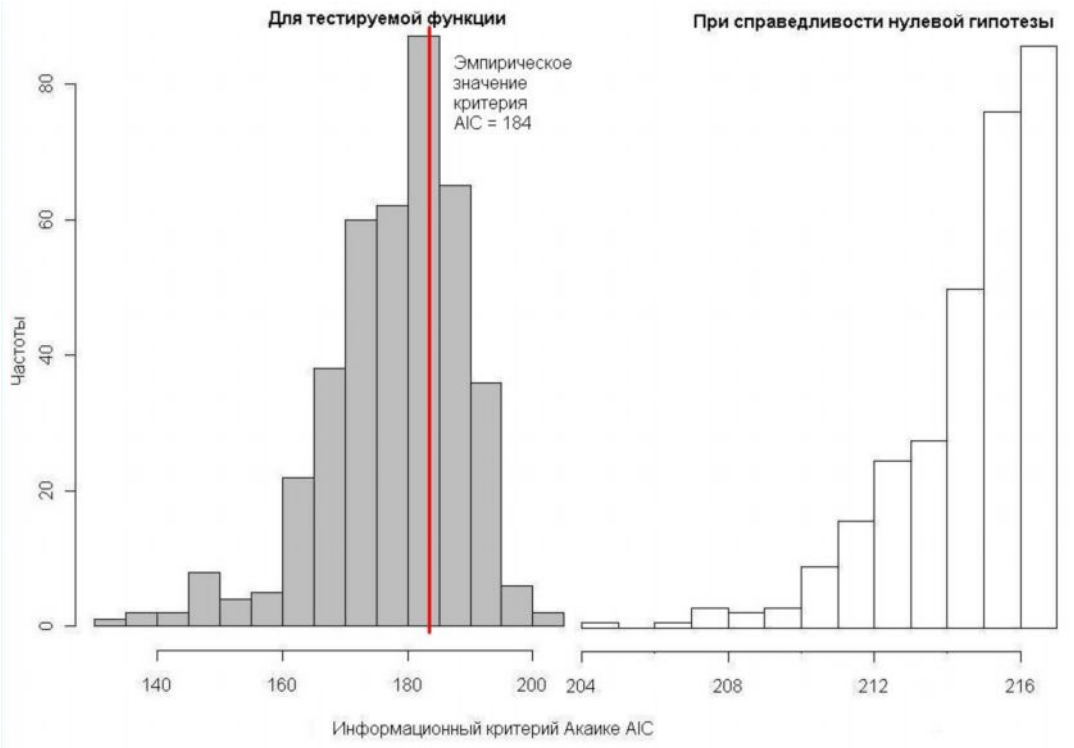


Рис.3.7. Гистограмма распределения AIC-критерия, полученная бутстрепом, для полинома 2-й степени (слева) и при справедливости нулевой гипотезы (справа).

Обратимся теперь к проблеме оценки доверительных интервалов для полученной модели. В общем случае различают три способа выражения интервальных оценок точности и воспроизводимости рассчитанной модели: (а) доверительная область модели регрессии, (б) доверительная область для значений прогнозируемой переменной и (в) толерантные интервалы регрессии (Кобзарь, 2006, с. 665).

Доверительный интервал модели регрессии (CI – *Confidence Interval*) для каждого текущего $x = x_0$ зависит от ошибки $SE(\hat{\mu}_{Y|x_0})$ расчетных значений $\hat{y}$, оцениваемых как $\hat{\mu}_{Y|x_0} = a + bx_0$. Для линейной регрессии одной переменной он рассчитывается как

$$\text{CI} = \left(a + bx_0 \pm t_{\frac{\alpha}{2},n-2} S_{\hat{y}} \sqrt{\frac{1}{n} + \frac{(\bar{x} - x_0)^2}{\sum (x_i - \bar{x})^2}} \right)$$, где $S_{\hat{y}}$ – стандартное отклонение остатков. (3.3)

Обычно это – довольно узкая полоса, которая несколько расширяется при крайних значениях *x*. Если многократно извлекать из генеральной совокупности наблюдений различные выборки из *n* пар (*x*, *y*) значений и строить по ним модели регрессии, то за пределы "доверительной полосы" может выйти только α% таких линий (обычно α = 5%).

Из этих соображений легко выполнить расчет CI бутстреп-методом (рис. 3.8):

а) выделяем *m* = 100 опорных значений, равномерно распределенных по шкале *x*, относительно которых будут рассчитываться величины доверительных интервалов;

б) делаем случайную выборку с возвращениями из порядковых номеров строк исходной таблицы и по этой перевыборке рассчитываем полиномиальную модель регрессии;

в) по модели п. б) выполняем расчет *m* опорных значений зависимой переменной;

г) пункты б-в) повторяем *B* = 1000 раз, после чего для каждой 100 опорных точек *x* вычисляем по 1000 расчетных значений $\hat{y}$, т.е. воспроизводим распределение прогноза отклика в этих точках;

д) для каждой из опорных точек *x* находим значения квантилей при *p* = 1 - α/2 и *p* = α/2 и вычисляем основные доверительные интервалы по формуле (1.9) раздела 1.4.

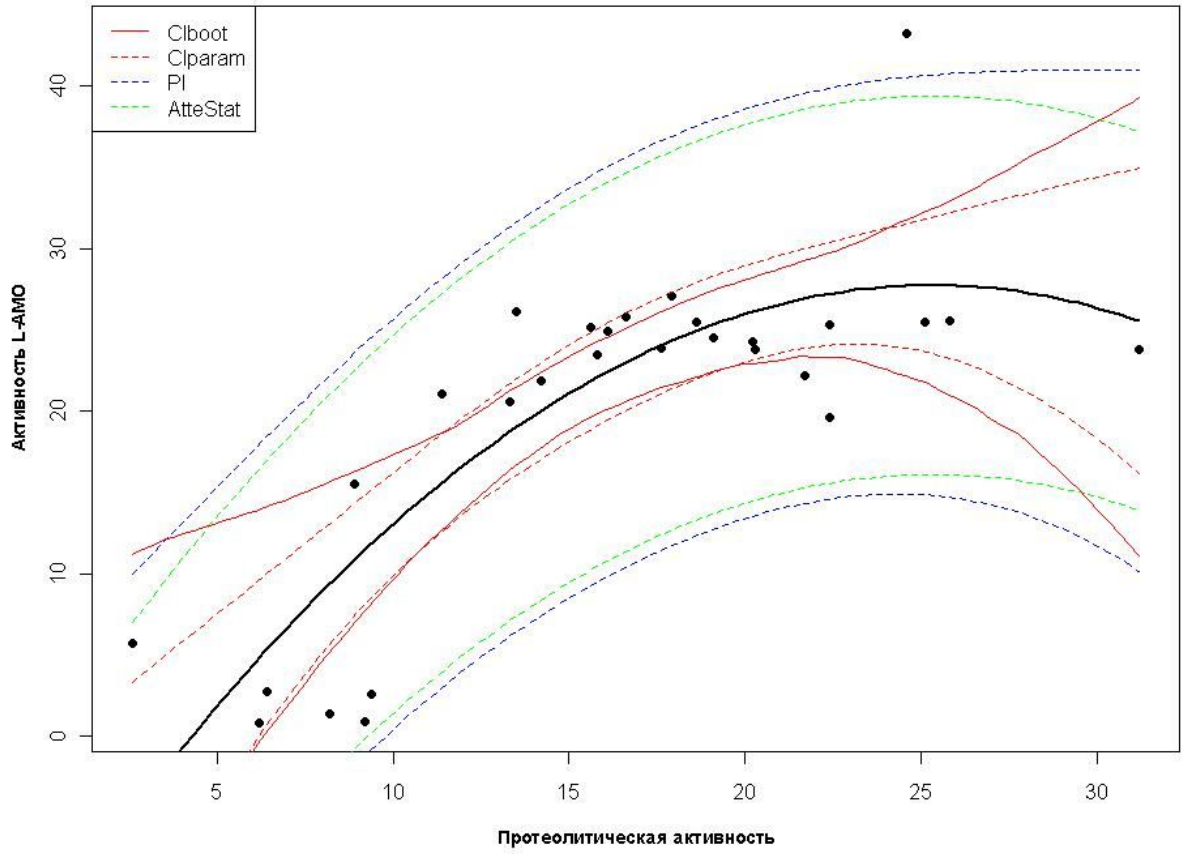


Рис.3.8. Различные формы 95% интервальных оценок полиномиальной модели: доверительные интервалы регрессии, полученные бутстреп-процедурой (CI_{boot}) и по параметрическим формулам (CI_{param}), интервалы предсказания (PI) и интервалы, рассчитываемые программой AtteStat

На графике, представленном на рис. 3.8, показано, что доверительные интервалы, найденные бутстреп-процедурой, в средней части шкалы *x* расположены в достаточной близости от границ, найденных по параметрическим формулам (3.3), однако существенно различаются в областях крайних значений *x*. Это связано с тем, что две точки с

"нехарактерными" величинами L-AMO своими отклонениями от регрессии нарушают нормальность распределения остатков, что делает использование параметрического метода проблематичным.

Доверительный интервал для зависимой переменной или интервал предсказания (PI – *Prediction Intervals*) складывается из разброса значений вокруг линии регресии и неопределенности положения самой этой линии. Его ширина определяется ошибкой $SE(\hat{y}\,|\,x_0) = SE(\hat{\mu}_{Y|x_0}) + \varepsilon_{Y|x_0}$, где $\varepsilon_{Y|x_0}$ – случайная условная ошибка при прогнозе *Y*. Для простой линейной регрессии его можно рассчитать по формуле (3.3), предварительно добавив 1 к подкоренному выражению. Считается, что интервал предсказания определяет диапазон разброса для "новых" прогнозируемых значений $\hat{y}$, хотя по его расчетной формуле это почти уникальное свойство предположить сложно.

В статистической среде R для линейной модели интервалы CI и PI легко различаются и рассчитываются с использованием функции predict.lm(…). Для других прикладных программ, что именно рассчитывается в качестве интервальных оценок, иногда бывает не вполне ясным. Например, в весьма достойной программе AtteStat в качестве "доверительных интервалов оценок модели" выводятся CI = ($\hat{y} \pm t_{\alpha/2,\ n-2}\ S_{\hat{y}}$), постоянные на всем диапазоне изменения *x* (см. рис. 3.8).

К разделу 3.4:

```
#  Определение векторов с активностями ферментов ядовитого секрета гадюк
x <- c(19.1, 22.4, 21.7, 20.2, 25.1, 18.6, 15.6, 17.6, 13.5, 22.4, 14.2, 13.3, 17.9, 24.6,
25.8, 16.1, 16.6, 31.2, 15.8, 11.4, 20.3, 6.2, 8.9, 9.4, 6.4, 9.21, 2.6, 8.2)
y <- c(24.5, 19.6, 22.2, 24.3, 25.5, 25.5, 25.2, 23.9, 26.1, 25.3, 21.9, 20.6, 27.1, 43.2,
25.6, 24.9, 25.8, 23.8, 23.5, 21.1, 23.8, 0.8, 15.5, 2.6, 2.8, 0.9, 5.7, 1.4)
xy <- data.frame(x,y)
library(compositions) ; library(boot)
# 1. Модели, линейные по параметрам
lmodel <- "y ~ x"   #  Линейная
#  Построение линейной модели с использованием функций lm(…) и glm(…)
fit <- lm(lmodel); summary(fit)
# Проверка гипотез - оценка статистической значимости R2 и AIC
# Определяем число итераций бутстрепа и эмпирические значения статистик
N = 1000; k1 = summary(fit)$r.square ; k2 = AIC(fit)
cnt1=0 ; cnt2=0  # Обнуляем счетчики числа ошибок 1-го рода
for(i in 1:N){ xp=sample(x) # перемешиваем значения x
       if (summary(lm(formula = y ~ xp))$r.square <= k1)  cnt2=cnt2 + 1
       if (AIC(lm(formula = y ~ xp)) >= k2)  cnt2=cnt2 + 1}
cnt1*100/N ; cnt2*100/N
# Оценка параметров бутстрепом – смещение среднего и доверительные интервалы AIC
vyb <- t(replicate(1000, sample.int(length(x), replace=TRUE)))
bAIC <- sapply(1:1000, function (i) {l<-vyb[i,]; AIC(lm(y[l] ~ x[l]))})
hist(bAIC); mean(bAIC) - k2 ; quantile(bAIC, prob=c(0.025,0.975))
fit <- glm(lmodel, gaussian); summary(fit)
#  Кросс-проверка модели по двум схемам разбиения: 1×28 и 4×7
cv.glm(xy, fit) ;  cv.glm(xy, fit, K = 7)$delta
#   Аналогичные вычисления проводим по моделям:
lmodel <- "y ~ I(1/x)"   #  Гипербола
lmodel <- " y ~ x + I(x*x)"   #  Полином 2 степени
lmodel <- " y ~ x + I(x*x)+ I(x*x*x)"   #  Полином 3 степени
#  Расчет различных вариантов интервальных оценок  (рис. 3.8)
m = 100 ; p2.fit <- lm( y ~ x + I(x*x)) ; xy <- data.frame(x,y); n <- nrow(xy)
sr <- qt(0.975, n-2)*summary(p2.fit)$sigma #  Стандартные отклонения для остатков
x.grid <- seq(from=min(x),to=max(x),length.out=m)
eval.grid <- data.frame(x=x.grid) # Сетка из m значений x
# Confidence Interval
p2.CI <- predict(p2.fit,newdata = eval.grid, level=0.95, interval="conf")
# Prediction Intervals
p2.PI <- predict(p2.fit,newdata = eval.grid, level=0.95, interval="pred")
#  Бутстреп-процедура расчета доверительных интервалов регрессии (3 связанных функции)
```

```
xy.resampler <- function() {   # Функция, выполняющая перевыборку исходной таблицы данных
       resample.rows <- sample(1:n,size=n,replace=TRUE); return(xy[resample.rows,]) }
xy.p2.estimator <- function(data,m=100) {  # Функция, рассчитывающая модель, по перевыборкам
       fit <- lm(y ~ x + I(x*x),data) ; return(predict(fit,newdata=eval.grid)) }
xy.p2.cis <- function(B,alpha,m=100) { #  Функция, рассчитывающая доверительные интервалы
       xy.main <- xy.p2.estimator(xy,m)
       xy.boots <- replicate(B,xy.p2.estimator(xy.resampler(),m))
       cis.lower <- 2*xy.main - apply(xy.boots,1,quantile,probs=1-alpha/2)
       cis.upper <- 2*xy.main - apply(xy.boots,1,quantile,probs=alpha/2)
       return(list(main.curve=xy.main,lower.ci=cis.lower,upper.ci=cis.upper)) }
xy.ci <- xy.p2.cis(B=1000,alpha=0.05)  #  Выполнение расчетов
#  Построение графика с различными вариантами интервальных оценок
plot(x, y, pch=16, main="") ; lines(x=x.grid, y=p2.CI[,1],lwd=2)
lines(x=x.grid, y=p2.CI[,2],col="red", lty=2); lines(x=x.grid, y=p2.CI[,3],col="red", lty=2)
lines(x=x.grid,y=xy.ci$lower.ci,col="red") ; lines(x=x.grid,y=xy.ci$upper.ci,col="red")
lines(x=x.grid, y=p2.PI[,2],col="blue", lty=2)
lines(x=x.grid, y=p2.PI[,3],col="blue", lty=2)
lines(x=x.grid, y=p2.CI[,1]+sr,col="green",lty=2)
lines(x=x.grid, y=p2.CI[,1]-sr,col="green",lty=2)
# 2. Нелинейные модели (тестируются последовательно)
fit <- nls(y ~exp(a*x)*x^b, start = list(a=0.3, b=1.1)); summary(fit)
#  Оценка значимости критерия R2
N = 1000; k1 = R2(fit)
vyr <- t(replicate(N, sample(length(x), replace=FALSE)))
bR2 <- sapply(1:N,
       function (i) {l<-vyr[i,]; R2(nls(y[l] ~exp(a*x)*x^b, start = list(a=0.3, b=1.1)))})
sum(bR2< k1)*100/N
#  Оценка доверительных интервалов коэффициентов регрессии
vyb <- t(replicate(1000, sample.int(length(x), replace=TRUE)))
bka <- sapply(1:1000, function (i) {l<-vyb[i,];
       fp <- coef (nls(y[l] ~exp(a*x[l])*x[l]^b, start = list(a=0.3, b=1.1))); fp[1]})
hist(bka);  quantile(bka, prob=c(0.025,0.975))   #  коэффициент a
bkb <- sapply(1:1000, function (i) {l<-vyb[i,];
       fp <- coef (nls(y[l] ~exp(a*x[l])*x[l]^b, start = list(a=0.3, b=1.1))); fp[2]})
hist(bkb);  quantile(bkb, prob=c(0.025,0.975)) #  коэффициент b
#  Кросс-проверка модели по схеме разбиения: 1×28
n <- length(x) ; rss=0
for(i in 1:n){xp = x[-i] ; yp = y[-i]
       c = coef(nls(yp ~exp(a*xp)*xp^b, start = list(a=0.3, b=1.1)))
       e = y[i] - (exp(c[1]*x[i])*x[i]^c[2])
       rss=rss + e*e} ; rss/n
#  Аналогично тестируются последовательно остальные нелинейные модели:
fit <- nls(y ~ a*exp(b*x), start = list(a=5, b=0.1)); summary(fit)
fit <- nls(y ~ a*x^b, start = list(a=0.3, b=1.1)); summary(fit)
fit <- nls(y ~a + b*x + c*log(x), start = list(a=0.3, b=1.1, c=1.1)); summary(fit)
fit <- nls(y ~a/(1+exp(b+c*x)), start = list(a=43, b=3, c=-0.15)); summary(fit)
```

3.5. Сравнение двух линий тренда и робастная регрессия

Характерную изменчивость значений популяционных показателей по шкале ведущего независимого фактора в экологии принято называть *градиентым трендом* (широтный градиент, высотный градиент, температурный градиент и др.). В простейшем случае можно предположить, что эта динамика имеет линейный характер. Рассмотрим, как в этих условиях можно проверить гипотезу о статистической идентичности скорости изменения изучаемого показателя для двух разных местообитаний.

В качестве примера сравним особенности высотного распределения дождевых червей семейства Lumbricidae в центральной части Северного Кавказа [П9]. Выборку из 149 средних значений популяционной плотности (численность и биомасса на м^2) для разных биотопов разделим на две группы: эльбрусский (1) и терский (2) варианты

поясности, различия между которыми определяются орографией районов. Для каждого из вариантов построим линии регрессии, вид которых представлен на рис. 3.9.

Предварительно выполним дисперсионный анализ общей линейной модели: $\ln(N) = \beta_0 + \beta_1 H + \beta_2 F + \beta_3 HF + \varepsilon$, где N – численность червей, H – высота биотопа, F – вариант поясности. Коэффициенты модели имели следующие статистические оценки:

Коэффициент	Значение	*t*-критерий	*p*-значение
β_1 (H)	-0.0005	-2.045	**0.0428**
β_2 (F)	-1.0406	-2.296	**0.0233**
β_3 (HF)	0.00042	1.441	0.152,

т.е. влияние обоих факторов (высоты и региона) оказалось статистически значимым.

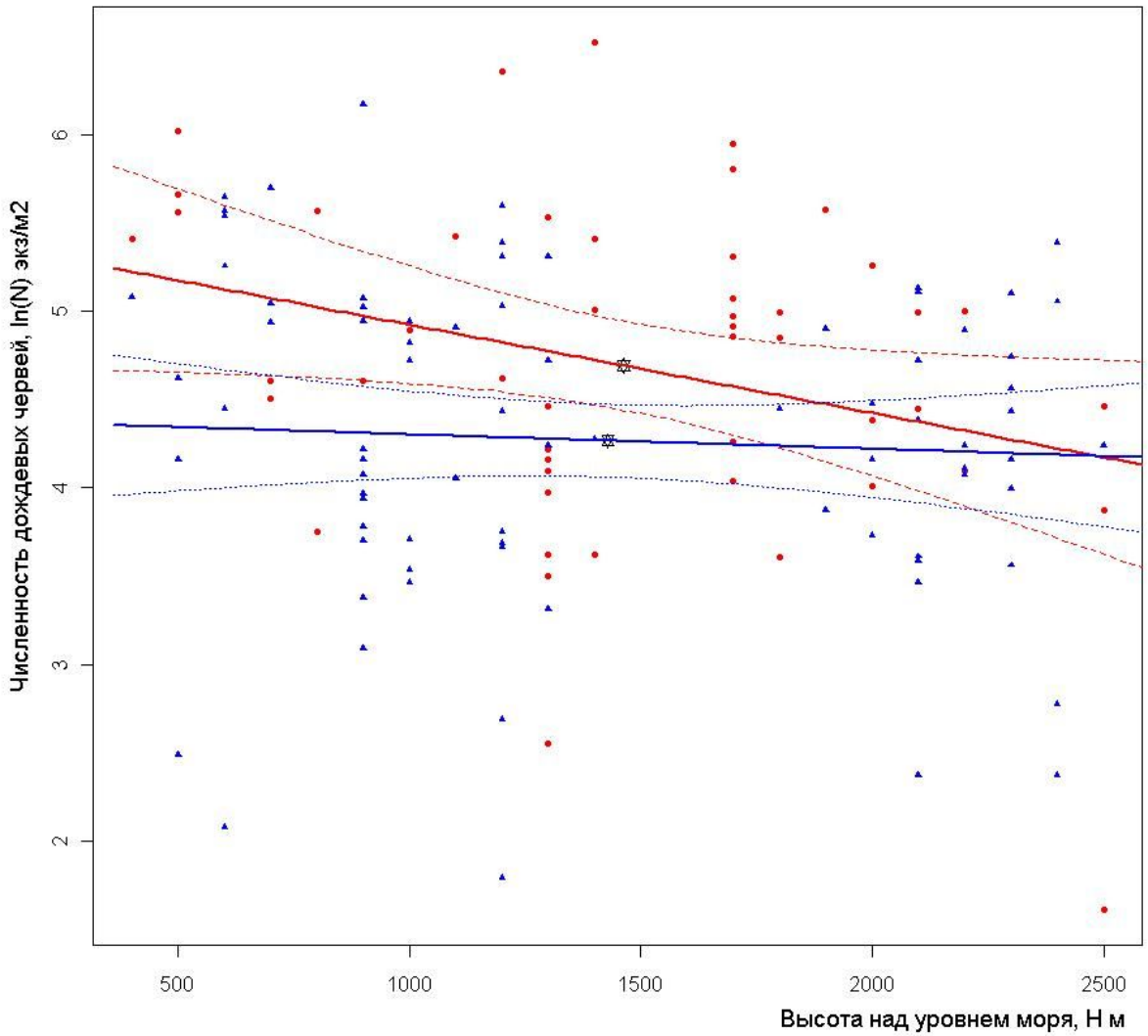


Рис.3.9. Линии регрессии и их доверительные интервалы (пунктир) зависимости популяционной плотности дождевых червей от высоты по эльбрусскому (красные цвета) и терскому (синие цвета) вариантов поясности

Важным моментом для нас является низкая оценка значимости коэффициента парного взаимодействия β_3 (HF), которая означает, что динамика снижения логарифмов численности популяций является статистически одинаковой для обоих вариантов поясности (в чем легко убедиться, проанализировав разложение на составляющие суммы квадраты отклонений для остатков). Дополнительно в этом можно убедиться, сравнив по критерию Фишера (F= 2.076, p = 0.152) дисперсии остатков для двух моделей – полной и без учета парного влияния факторов $\ln(N) = \beta_0 + \beta_1 H + \beta_2 F + \varepsilon$.

В практической плоскости обычно ставится вопрос не о полном *совпадении* трендов, а о *параллельности* двух линий регрессии $y_1 = a_1 + b_1x$ и $y_2 = a_2 + b_2x$, т.е. проверяется гипотеза H_0: $b_1 = b_2$. Для этого можно воспользоваться вполне тривиальным способом: рассчитать статистику Стьюдента $t = (b_1 - b_2)/s$ и найти соответствующее ей *p*-значение. Если в дальнейшем ставится задача оценить уровень полного совпадения моделей, то должна быть проверена и вторая гипотеза H_0: $a_1 = a_2$.(детали см. в Good, 2006б).

Поскольку стандартный регрессионный анализ изначально предполагает оценку ошибок коэффициентов s_1 и s_2, то методологическая трудность лишь в решении проблемы Беренса-Фишера: следует ли усреднять дисперсии по методу Уэлча (Аспина, Сатервайта и др.) или придумать какую-либо свою приближенную формулу $s = f(s_1, s_2)$. Разумеется, всех этих трудностей можно избежать, если применить рандомизационный тест, состоящий из следующих действий:

° многократно (*B* раз) перемешиваем метки группировочного фактора *F* относительно строк *x*-*y* исходной таблицы;

° для каждой пары случайно сформированных выборок рассчитываем уравнения регрессии и находим разность коэффициентов угла наклона ($b_1^* - b_2^*$);

° подсчитываем число случаев *k*, когда $|b_1^* - b_2^*|$ превысило эмпирическую разность $|b_1 - b_2|$, а долю k/B от общего числа итераций будем интерпретировать как статистическую значимость *p* нулевой гипотезы H_0: $b_1 = b_2$.

Результаты выполнения теста на параллельность линий регрессии с использованием параметрического и рандомизационного методов представлены в табл. 3.5. Они полностью соответствуют выводам дисперсионного анализа о том, что между двумя горными массивами нет статистически значимых различий в динамике влияния высотного фактора на популяции дождевых червей. К сожалению, строгость этого заключения является неполной, поскольку для терского варианта поясности коэффициенты угла наклона b_2 сами являются статистически незначимыми.

Таблица 3.5. Статистический анализ коэффициентов угла наклона *b* линий высотного (*H*) тренда численности популяций дождевых червей для эльбрусского и терского вариантов поясности; *N* – численность, *B* – биомасса, *t*-крит – приведенная разность *b* для разных вариантов, p_{par}, p_{ran} – *p*-значения для H_0: $b_1 = b_2$, рассчитанные параметрическим методом и рандомизацией соответственно, τ - ранговая корреляция Кендалла; жирным шрифтом отмечены статистически значимые модели

Модель регрессии	Вариант	Классическая регрессия МНК				Регрессия Кендалла-Тейла		
		b	Pr(*b* = 0)	*t*-крит	p_{par}/p_{ran}	*b*	τ	Pr(τ = 0)
$\ln(N) = a + bH$	Эльб.	**-0.0005**	**0.038**	1.474	0.142 / 0.149	-0.00036	-0.150	0.122
	Терс.	-0.00008	0.610			-0.00014	-0.061	0.405
$B = a + bH$	Эльб.	-0.0176	0.184	-0.667	0.505 / 0.456	-0.0137	**-0.200**	**0.038**
	Терс.	**-0.031**	**0.050**			-0.006	-0.067	0.361
$\ln(B) = a + bH$	Эльб.	**-0.00064**	**0.021**	1.48	0.141 / 0.166	-0.00053	**-0.200**	**0.038**
	Терс.	-0.00013	0.523			-0.00018	-0.067	0.361
$(NB)^{0.5} = a + bH$	Эльб.	-0.0307	0.138	0.722	0.471 / 0.365	-0.023	-0.156	0.107
	Терс.	-0.0143	0.138			-0.0097	-0.075	0.311
$\ln(NB)^{0.5} = a + bH$	Эльб.	**-0.00057**	**0.019**	1.63	0.105 / 0.134	-0.00042	-0.156	0.107
	Терс.	-0.00011	0.496			-0.00018	-0.075	0.311

Выполненные расчеты подтверждают основной постулат классического регрессионного анализа: корректность результатов напрямую зависит от характера распределения ошибок, т.е. остатков модели ε. В частности, удачно проведенное логарифмирование или иное нелинейное преобразование данных может привести к полному изменению статистических выводов: см. эффект "перевертыша" значимости моделей по биомассе в табл. 3.5.

Некоторых подобных недостатков лишены альтернативные непараметрические методы, такие как робастная линейная регрессия Кендалла-Тейла (KTR – Kendall-Theil robust line, см. Helsel, Hirsch, 2002), которая имеет вид $y = \hat{\beta}_0 + \hat{\beta}_1 x + \varepsilon$. Крышечка ^ над символами β означает, что коэффициенты регрессии являются не параметрами некой теоретической модели, а некоторой статистикой на множестве возможных значений. Регрессия Кендалла-Тейла тесно связана с коэффициентом ранговой корреляции τ Кендалла, т.е. проверка гипотезы $H_0: \hat{\beta}_1 = 0$ идентична оценке значимости H_0: τ = 0.

Основные процедуры подбора уравнения регрессии KTR сводятся к следующему:

° вычисляется множество $n(n-1)/2$ коэффициентов угла наклона $b_{ij} = (y_i - y_j)/(x_i - x_j)$ для всех возможных пар точек i и j исходной выборки;

° расчетные коэффициенты модели оцениваются как медианы

$$b = \text{med}(b_{ij}); \quad a = \text{med}(y) - b \cdot \text{med}(x);$$

° доверительные интервалы коэффициентов рассчитываются по интерполяционным формулам (см. работу Helsel, Hirsch, 2002 или прилагаемый скрипт Kendall_Theil_Regr.r, разработанный М. Вейнкауфом), а статистическая значимость модели в целом оценивается по формулам для коэффициента ранговой корреляции τ Кендалла.

Поскольку доверительные интервалы коэффициентов b для всех пяти моделей регрессии KTR пересекаются, нет оснований отклонять нулевую гипотезу об одинаковой изменчивости популяционной плотности дождевых червей по шкале высотного градиента для рассматриваемых горных массивов. Обратим также внимание на то, что коэффициент ранговой корреляции τ не зависит от способа нелинейного преобразования данных – см. табл. 3.5. В то же время мы напоминаем о низкой мощности непараметрических тестов и воздерживаемся от категоричных рекомендаций по их безоговорочному применению.

К разделу 3.5:

```
LUM<-read.table("lumbr.txt",header=TRUE,sep="\t") ; attach(LUM)
# Дисперсионный анализ линейных моделей с учетом двух факторов
dat <- data.frame(x=H, y=log(CHISL), f=factor(VARIANT))   #  отклик ln от численности
summary(lm(y~x*f,data=dat)) ; anova(lm(y~x+f,data=dat),lm(y~x*f,data=dat))
#  Функция расчета t-стат. при сравнении b1 и b2  и  p-значения аппроксимацией
test1 <- function(dat, fac=dat$f) {   fits <- lapply(split(dat,fac),lm,formula=y~x)
       sums <- lapply(fits,summary) ;   coefs <- lapply(sums,coef)
       db <- coefs[[2]]["x","Estimate"]-coefs[[1]]["x","Estimate"]
       sd <- sqrt(sum(sapply(coefs,function(x) x["x","Std. Error"])^2))
       df <- sum(sapply(fits,"[[","df.residual")) ;   td <- db/sd
       c(est=db, sd=sd, tstat=td, prt=2*pt(-abs(td),df))  }
#  Те же результаты можно получить, оценив парное взаимодействие двух факторов
test2 <- function(dat) {   fit <- lm(y~x*f,data=dat);   coef(summary(fit))["x:f2",]}
rbind(test1(dat),test2(dat))   #  Небольшая разница за счет неточности определения df
source("print_rezult.r") #  Функция, выполняющая пермутационный тест и расчет p-значения
testPerm <-  function(dat, Nperm=1000)  {
  PermArray <- as.numeric(rep(NA, Nperm)) ;  for (i in 1:Nperm) #  перемешиваем только фактор
  PermArray[i] <- test1(dat,sample(dat$f))[3] ;
  return (RandRes(test1(dat)[3], PermArray, Nperm)) }
testPerm(dat)
#  Расчеты повторяем, используя в качестве отклика разные показатели популяционной плотности
dat <- data.frame(x=H, y= BM, f=factor(VARIANT))   #  отклик  -  биомасса
dat <- data.frame(x=H, y=log(BM), f=factor(VARIANT))   #  ln от биомассы
dat <- data.frame(x=H, y=sqrt(CHISL* BM), f=factor(VARIANT))   #  индекс Гилярова
dat <- data.frame(x=H, y=log(sqrt(CHISL* BM), f=factor(VARIANT)))   # ln от индекс Гилярова
#   ------------------------------------------------------------
#  Вывод графиков линий регрессии и их доверительных интервалов
df1 <- data.frame( x=H[VARIANT==1], y=log(CHISL[VARIANT==1])) # Другой вариант
df2 <- data.frame(x=H[VARIANT==2], y=log(CHISL[VARIANT==2]))  # исходных таблиц
xin <- seq(0.9*min(H), 1.1*max(H), length=100) ; fit1 <- lm(y~x, df1); fit2 <- lm(y~x, df2)
pre1 <- predict(fit1, data.frame(x=xin), interval="confidence")
pre2 <- predict(fit2, data.frame(x=xin), interval="confidence")
```

```
pre <- data.frame(pre1, pre2)
plot(df1, pch=16, cex=0.7, col="red"); points(mean(df1$x), mean(df1$y), pch=11)
points(df2, pch=17, cex=0.7, col="blue"); points(mean(df2$x), mean(df2$y), pch=11)
matplot(xin,pre,type="l",lty=c(1,2,2,1,3,3),
                    lwd=c(2,1,1,2,1,1), col= c(2,2,2,4,4,4), add=TRUE)
source("Kendall_Theil_Regr.r")    # Расчеты характеристик регрессии  Кендалла-Тейла
(Res1 <- Kendall(df1)) ; (Res2 <- Kendall(df2))
plot(df1,type="p",cex=0.7, col="red",pch=16) ; points(df2, pch=17, cex=0.7, col="blue")
curve(Res1[1,1]*x+Res1[4,1], add=TRUE, col="red", lwd=2)
curve(Res2[1,1]*x+Res2[4,1], add=TRUE, col="blue", lwd=2)
#  Расчеты повторяем, используя в качестве отклика разные показатели популяционной плотности
```

3.6. Модели распределения популяционной плотности по градиенту

Реакция экологических сообществ на изменение условий среды, выражающаяся в изменении видового состава и плотности популяций, зависит от всего комплекса воздействующих факторов, под которыми понимается любая внешняя или внутренняя движущая сила, модифицирующая показатели жизнедеятельности экосистемы. Количественная оценка экологического отклика различных видов живых организмов на внешние и внутренние возмущения является основной задачей моделирования окружающей среды (habitat models).

Традиционно используемые статистические методы предполагают, что кривая зависимости популяционной плотности y от величины воздействующего фактора x имеет симметричную форму колоколообразной гауссианы $y = he^{-(x-\mu)^2/2\sigma^2}$ с тремя интерпретируемыми параметрами: μ – локальный оптимум на оси x, которому соответствует максимум обилия вида h; σ – стандартное отклонение на шкале градиента относительно этого оптимума.

Однако в результате сложного воздействия коррелированного комплекса факторов, эффект влияния которых трудно идентифицировать в отдельности, эта зависимость часто имеет ярко выраженный ассиметричный или полимодальный характер. Поэтому, сохранив традиционную симметричную гауссову модель только как образец для сравнения, были разработаны статистические методы аппроксимации асимметричных кривых отклика различными типами моделей (Guisan, Thuillier, 2005; Шитиков и др., 2011):

- обобщенными линейными моделями GLM (generalized linear model);
- обобщенными аддитивными моделями GAM (generalized additive model);
- моделями HOF Хаусмана-Олфа-Фреско (Huisman et al., 1993);
- обобщенными линейными моделями пространственного распределения (GRASP) и моделями сверхниши (hyper niche).

Модели GLM и GAM являются частным случаем обобщенного уравнения нелинейной регрессии

$$y = g^{-1}\left[\sum_{i=1}^{n} a_i q_i(x_i)\right], \tag{3.4}$$

где $g(y^{-1})$ является *функцией связи* (link function), а $q_i(x_i)$ – произвольными функциями нелинейного преобразования независимых факторов. Если принимаются некоторые условия тождественности этих функций, то мы получаем следующие классы моделей:

- при $q_i(x_i) = x_i$ – обобщенную линейную модель $y = g^{-1}\left(\sum_{i=1}^{n} a_i x_i\right)$;
- при $g(y') = y$ – обобщенную аддитивную модель $y = \sum_{i=1}^{n} a_i q_i(x_i)$;
- при $g(y') = y$ и $q_i(x_i) = x_i$ – обычную линейную регрессию $y = \sum_{i=1}^{n} a_i x_i$.

Обобщенные линейные модели (McCullagh, Nelder, 1989; Venables, Ripley, 2002) дают возможность использовать различные функции связи, конкретный выбор которых зависит обычно от природы случайного распределения отклика y и его остатков. Если мы располагаем выборочными значениями переменных y и x, измеренными в непрерывных

шкалах, и нет оснований отказываться от предположения о нормальном распределении остатков, то функция преобразования задается идентичностью $q(x) = x$.

Отклик y может выражаться дихотомической переменной, связанной только, например, с фактом встречаемости вида. Если p – доля единичных значений y, а $(1 - p)$ – доля нулевых значений для того же вида, то функция связи обычно задается в виде логита (logit link) или логарифма отношения шансов "встретить/не встретить особь данного вида": $g(y) = \log(\frac{y}{1-y}) = \log(\frac{p}{1-p})$. Соответствующая ей обобщенная логит-линейная модель $\log \frac{p(y)}{1-p(y)} = \sum_{i=1}^{n} a_i x_i$ характеризуется кривой S-образной формы, а ее параметры можно интерпретировать следующим образом: при изменении значения предиктора на единицу, значение логарифма отношения шансов зависимой переменной изменится на величину соответствующего коэффициента. Логистическая модель предполагает биномиальный закон распределения ошибок, что делает некорректным использование стандартного метода наименьших квадратов.

Если отклик y равен количеству независимых событий, произошедших с постоянной скоростью за некоторый промежуток времени, а предикторы – категориальные переменные, то мы должны воспользоваться моделью регрессии Пуассона: $\log(y) = \mu + \lambda^A + \lambda^B + \lambda^{AB}$ (для двух факторов A и B).
Здесь $g(y) = \log(y)$ – функция связи, λ – дифференциальные эффекты факторов.

Интересно проследить, например, математическую идентичность гауссовой модели отклика и GLM в форме полинома 2-го порядка при логарифмической функции связи

$$\log(y) = a_0 + a_1 x + a_2 x^2 \quad \Leftrightarrow \quad y = h \exp[-(x-\mu)^2 / 2\sigma^2],$$

если удовлетворяются условия $\mu = -a_1/2a_2$; $\sigma = (-1/2a_2)^{0,5}$; $h = \exp(a_0 - a_1^2/4a_2)$. Использование иных коэффициентов или полиномов 3-го и более порядков будет подчеркивать отличия асимметрических отклонений функции отклика от гауссианы.

Таким образом, обобщенные линейные модели (GLM) расширяют диапазон приложений линейного моделирования, рассматривая широкий класс альтернативных распределений в случае неустойчивых дисперсий или нарушений предположений об их нормальности.

Разумеется, использование обычного метода наименьших квадратов для настройки параметров моделей GLM и GAM является некорректным. К асимптотически (т.е. при $n \to \infty$) оптимальным значениям неизвестных параметров для широкого класса вероятностных моделей приводит *метод максимального правдоподобия* (MLE – Maximum Likelihood Estimation). В случае достаточных статистик он превосходит метод моментов и дает наилучшие оценки с точки зрения квадратичного риска. Принцип максимального правдоподобия состоит в том, что в качестве "наиболее правдоподобного" значения параметра берут значение Θ, максимизирующее вероятность получить при n опытах имеющуюся выборку $X = (x_1, \ldots x_n)$.

Предположим, что нам известен закон распределения дискретной случайной величины X, определяемый параметром Θ, но неизвестно численное значение этого параметра. И пусть $p(x_i, \Theta)$ – вероятность того, что в результате испытания величина X примет значение x_i. *Функцией правдоподобия* случайной величины X называют функцию аргумента Θ, определяемую по формуле:

$$L(x_1, x_2, \ldots, x_n; \Theta) = p(x_1,\Theta)p(x_2,\Theta)\ldots p(x_n,\Theta).$$

И тогда в качестве искомой точечной оценки параметра Θ принимают такое его значение, при котором функция правдоподобия достигает максимума: $\theta_{MLE} = \arg\max L(\mathbf{x}|\theta)$, $\theta \in \Theta$.

Для поиска экстремума необходимо найти частную производную функции правдоподобия и приравнять ее нулю: $\partial L(\mathbf{x}, \theta_o)/\partial\theta = 0$. Если вторая производная в критической точке отрицательна, то это – точка максимума, соответствующая искомому значению параметра. Если нам необходимо найти k параметров, то формируется и

решается система из k уравнений правдоподобия. Поскольку функции L и $\ln L$ достигают максимума при одном и том же значении Θ, удобнее искать максимум логарифмической функции правдоподобия $\ln L$.

Например, в стандартном методе наименьших квадратов (МНК) ищется минимум квадратов разности между эмпирическими и полученными по модели значениями отклика $\sum_{i=1}^{n}(y_i - \mu_i)^2$. Чтобы найти наиболее правдоподобные значения μ_i, необходимо проанализировать функцию максимального правдоподобия L. Если предположить, что функция плотности распределения каждого y_i подчиняется нормальному закону, а параметр дисперсии σ постоянен для всех наблюдений, то:

$$L(\mu, y) = \sum_{i=1}^{n}\{-(y_i - \mu_i)^2 / 2\sigma^2 - \ln(\sqrt{2\pi\sigma}\},$$

а наиболее правдоподобные MLE-оценки μ_i будут идентичны полученным МНК.

Если мы проводим регрессионный анализ в рамках предположений классической модели, то можно записать функцию правдоподобия в терминах остатков:

$$L(\boldsymbol{\varepsilon} \mid \boldsymbol{\theta}; \sigma^2) = -n\ln(2\pi)/2 - n\ln(\sigma^2)/2 - [(\mathbf{Y} - \mathbf{X\theta})^T(\mathbf{Y} - \mathbf{X\theta})]/2\sigma^2.$$

Однако другие функции вероятности распределения остатков (например, функция Пуассона $e^{-\mu_i}\mu_i^{\,y_i} / y_i!$) приводят к другим выражениям для функции максимального правдоподобия. Тогда, чтобы найти искомые параметры μ_i, а, следовательно, и коэффициенты GLM-моделей a_0, a_1, a_2,… необходимо воспользоваться итеративными нелинейными процедурами оптимизации функции L (Oberhofer**,** Kmenta, 1974; Oksanen, Minchin, 2002).

Рассмотрим модели отклика на примере оценки вероятности встречаемости отдельных видов и подсемейств макрозообентоса на различных участках р. Сок [П2]. Ранее отмечалось, что упорядоченную последовательность **X** значений произвольного независимого фактора в экологии принято называть *градиентом*. Однако непонятно, какой конкретно показатель из многих возможных следует считать продольным градиентом реки: расстояние до устья, высоту над уровнем моря, ширину, скорость течения, температуру воды либо что-то еще?

Распространенные в 70-е годы XX в. концепции *прямого градиентного анализа*, учитывающие только о д и н параметр среды, показали свою несостоятельность в сколько-нибудь сложных региональных условиях. Выход может быть найден в редукции исходного множества переменных и сведении их к небольшому количеству латентных факторов, обладающих некоторым внутренним единством и соотнесенных с определенным биологическим смыслом. Для этого воспользуемся представленным далее в главе 6 методом *главных компонент*, формирующим в пространстве признаков набор новых ортогональных осей, которые проходят сквозь эллипсоид облака точек в оптимальных направлениях максимальной изменчивости.

Две главные компоненты (рис. 3.10), обобщающие 12 различных геофизических и гидрохимических показателей, измеренных на водотоке р. Сок вместе с его притоком р. Байтуган [пример П2], объясняют 68,3% общей изменчивости этих признаков. Первый фактор (49.4% объясненной вариации) с большой очевидностью связан с пространственным градиентом и объединяет высоту над уровнем моря, глубину и температуру воды в месте отбора проб, изменчивость грунтов и т. д. Второй фактор связан с оценкой качества воды и объединяет органическое загрязнение, насыщенность воды кислородом, а также площадь водосбора, как источник аккумуляции органического вещества. Для удобства последующей интерпретации примем шкалу комплексного градиента в виде последовательности значений $x = F_{\max} - F = 1.28 - F$, где F – фактор 1 на рис. 3.8, т.е. величина градиента почти монотонно увеличивается от 0 (исток р. Байтуган) до 2.81 (ст. 12 на приустьевом участке р. Сок).

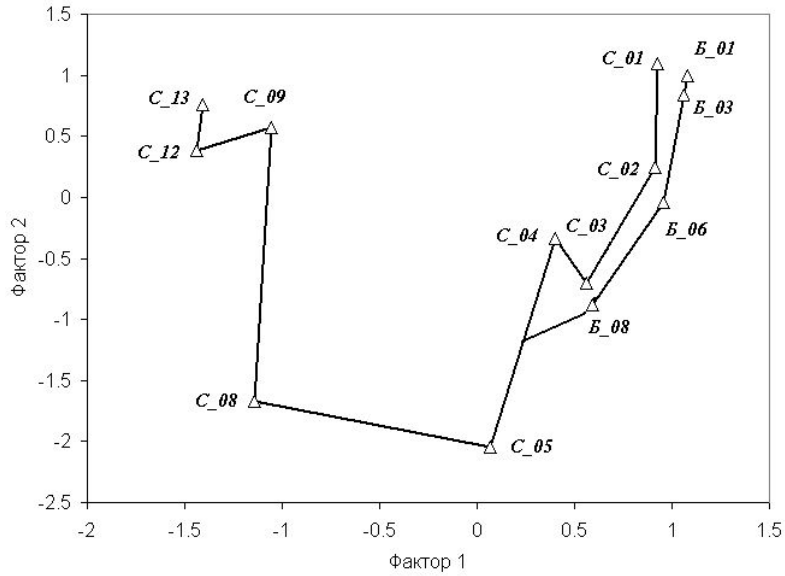


Рис. 3.10. Диаграмма расположения участков рек Сок (префикс "С") и Байтуган (префикс "Б") в пространстве двух главных компонент и относительные нагрузки исходных переменных, вносимые в эти факторы

Показатель	Фактор 1	Фактор 2
Доля песчано-гравийных субстратов	**0,154**	0,027
Высота над уровнем моря	**0,149**	0,116
Азот нитритный	**-0,132**	0,046
Глубина отбора проб	**-0,148**	0,038
Доля илистых грунтов	**-0,149**	-0,034
Температура	**-0,157**	0,041
pH дна	-0,112	-0,018
Минеральный фосфор	-0,123	**-0,226**
Скорость течения	0,078	**-0,312**
Площадь водосбора	0,014	**-0,332**
Бихроматная окисляемость	-0,016	**-0,291**
Содержание кислорода	-0,050	**0,276**

Выше отмечалось, что одно из главных предположений градиентного анализа, традиционно ориентировавшегося на геоботанические представления, – гауссовый характер зависимости обилия вида от фактора среды. Для его проверки выполним с использованием различных моделей GAM, GML и HOF оценку распределения встречаемости особей отдельных таксонов макрозообентоса по шкале комплексного градиента x рек Байтуган-Сок. Расчеты будем выполнять с использованием модуля R "Species response curves", разработанного Д. Зелены (Zelený) в рамках надстройки к программе JUICE (Tichý , 2002) – подробности см. в тексте скрипта к этому разделу.

Логит-линейная модель (GML 1-го порядка), оценивающая вероятность обнаружить при текущем значении градиента x хотя бы один вид из подсемейства Ortocladeiinae имеет формулу:

$$g(y) = a_0 + a_1 x = 1.17 - 8.18\, x,$$

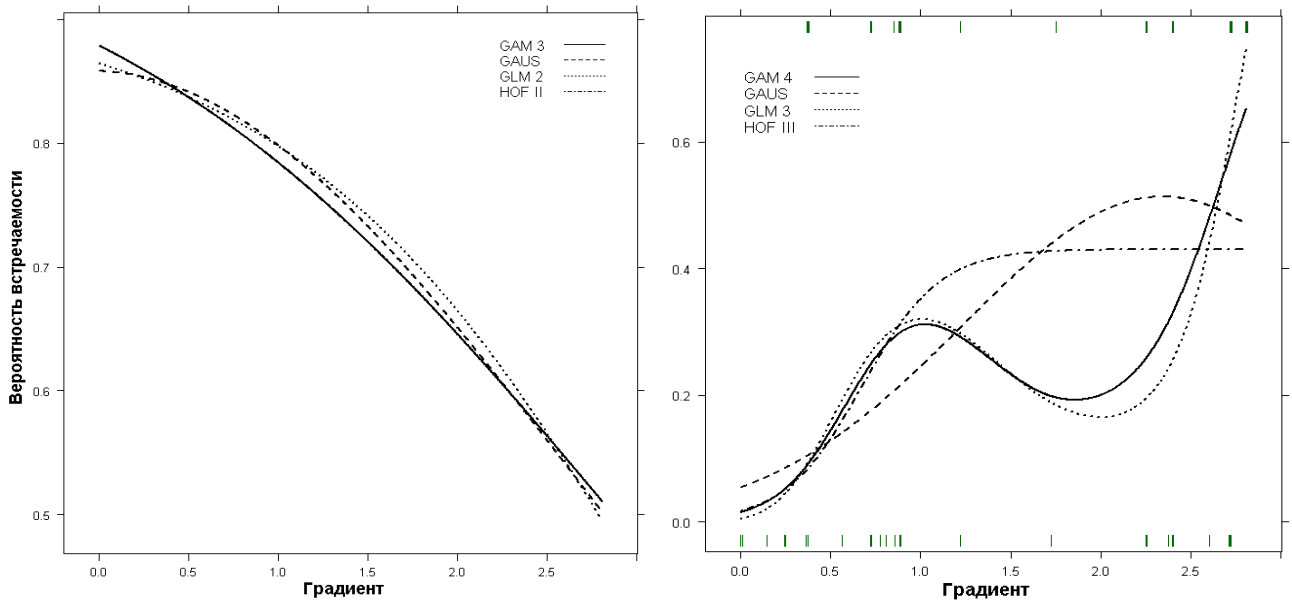
т.е. шанс встретить эти организмы монотонно уменьшается при движении от истоков к устью. Статистическая значимость коэффициента a_1, определяющего эту зависимость, весьма велика: критерий $z = -3.517$ при $p = 0.000437$. Использование рандомизационного теста, проверяющего нулевую гипотезу H_0: $|a_1| = 0$, после 1000 итераций привело к похожим результатам: $p = 0.001$.

Аналогичная модель GML 2-го порядка имеет вид

$$g(y) = a_0 + a_1 x + a_2 x^2 = 1.17 - 8.04\, x - 0.8\, x^2 ,$$

однако коэффициент a_2 оказался уже статистически незначим ($p = 0.732$ по рандомизационному тесту), а оценка информационной избыточности по критерию Акаике при переходе к нелинейной модели увеличилась со 170 до 175.

Остальные модели, такие как гауссова, GAM и HOF, в отношении видов Ortocladeiinae проявили очевидную синхронность в выводах (см. рис. 3.11а), однако такое единодушие встречалось не всегда. Например, для *Cladotanytarsus mancus* общая интерпретация осталась неясной, поскольку, в отличие от модели HOF, обобщенные модели GLM на основе полинома 3-й степени и GAM с 4-мя узлами сглаживания уверенно демонстрируют полубимодальную кривую (рис. 3.11б).



а) Ortocladeiinae б) *Cladotanytarsus mancus*

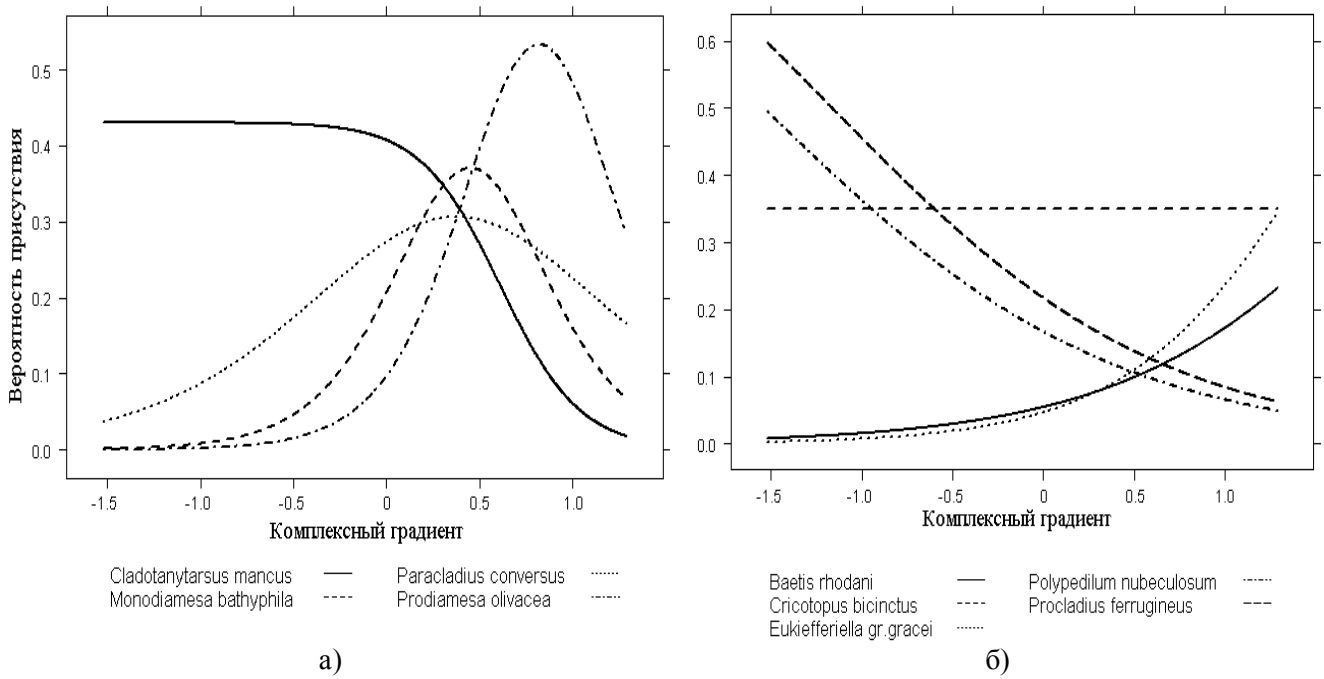
Рис. 3.11. Оценка распределения вероятности встречаемости таксонов макрозообентоса на шкале комплексного градиента с использованием различных моделей отклика

Какая же из моделей предпочтительней? И здесь уместно подробнее представить модель HOF (Huisman et al., 1993), которая позволяет гибко учитывать всю совокупность априорных исходных ограничений и теоретических предположений, традиционно связываемых с характером кривых отклика, и, вероятно, предоставляет наилучший результат с экологической точки зрения. Эта модель может быть в общем виде выражена, как $y(a, b, c, d, x, M) = \frac{M}{[1+\exp(a+bx)][1+\exp(c-dx)]}$, т. е. отклик y зависит от значения градиента x, максимально возможного значения M и четырех параметров $\{a, b, c, d\}$. Каждый из параметров может быть свободно варьируемым или равным определенному фиксированному значению и, в соответствии с этим, модель может принимать иерархическое множество состояний из пяти возможных форм (I–V):

Формы модели HOF	Список параметров				Кол-во
• V – асимметричный унимодальный отклик	***a***	***b***	***c***	***d***	4
• IV – симметричный унимодальный отклик	***a***	***b***	***c***	*b*	3
• III – монотонный рост с "плато"	***a***	***b***	***c***	∞	3
• II – монотонный рост	***a***	***b***	0	0	2
• I – "плато" (отсутствие отклика)	***a***	0	0	0	1

Самая сложная форма модели HOF – асимметричная кривая, включающая полный комплект из всех четырех параметров. Иерархичность здесь заключается в том, что более простая модель может быть получена из более сложной модели путем фиксации значений одного из параметров (Oksanen, Minchin, 2002). Выбор одной из пяти возможных форм осуществляется с использованием достаточно сложной автоматической процедуры на основе оценок стандартных отклонений и информационных критериев AIC или BIC.

Модели HOF проводят поиск только унимодальных зависимостей (рис. 3.12), которые необходимы для нахождения оптимума вида на градиенте, что обеспечивает их "шумозащищенность". Модели GLM и GAM в таких сложных ситуациях стремятся увеличить число степеней свободы и могут легко придти к "полубимодальной" (semi-bimodal) зависимости с неясной содержательной интерпретацией (рис. 3.11б).



а) б)

Рис. 3.12 Распределение обилия некоторых видов макрозообентоса в гидробиологических пробах по шкале комплексного градиента с использованием модели HOF

Таким образом, классическая унимодальная форма кривой отклика с максимумом в середине шкалы продольного градиента *x* оказалась характерной для немногочисленного и мало интересного с точки зрения биоиндикации подмножества видов макрозообентоса с широким диапазоном толерантности к абиотическим факторам. Возможно, что размах градиента оказался в нашем случае недостаточно широк для некоторых видов, либо выборка была недостаточно репрезентативная... Впрочем, более подробное обсуждение этих проблем уже выходит за рамки тематики нашей книги.

К разделу 3.6:

Предварительные операции:

а) Установить пакеты tcltk, lattice, mgcv и gravy. При использовании R версии 2.12 и выше пакет gravy следует устанавливать из источника http://sci.muni.cz/botany/zeleny/R/windows/contrib/R-2.12/

б) Загрузить из ресурса http://www.sci.muni.cz/botany/zeleny/wiki/juice-r/doku.php?id=scripts скрипт species_response_curves.r

в) Подготовить в файлах данные для трех таблиц:

- JUICE.short.head со столбцами Releve_Number (порядковые или идентификационные номера биотопов); Group_Number (произвольный номер группы биотопов, например, 1); Short_Head (значение показателя, соответствующего градиенту);
- JUICE.species.data со столбцами SpecName_Layer (сокращенное наименование вида); Species_Name (полное наименование вида); Layer (слой биотопа, например, 1);
- JUICE.table, в каждой строке которой перечислены номер биотопа (Releve_Number) и значения обилия видов; наименования столбцов должны соответствовать значениям полей SpecName_Layer таблицы JUICE.species.data.

```
#  Запуск скрипта диалогового построения моделей отклика компонентов экосистемы
  JUICE.short.head <- read.table("ShortHead.txt", sep="\t", head = T)
  JUICE.species.data <- read.table ("SpeciesData.txt", sep="\t",head = T)
  JUICE.table <- read.table ("table.txt", sep = "\t", check.names = F, head = T, row.names =
1)
source("species_response_curves.r")
#  Результаты представлены в графическом окне и файле result_table.csv
#   Если полученных результатов окажется недостаточно, то необходимо в нужных точках скрипта
#    расставить команды сохранения в файле объектов, созданных программой.
#   Например, в теле функции count.GLM в строке 61 скрипта species_response_curves.r вставим
```

```
#    команду  save(glm.1,glm.2,part.species, gr, file = "xy.RData")
#   Далее продолжить анализ с использованием сохраненных объектов:
load("xy.RData")
#  Вывод информации о моделях GLM  1-го и второго порядка
summary(glm.1) ; summary(glm.2)
Nrand = 1000  #  Число перевыборок рандомизации
#  Получение 1000 выборок со случайно переставленными значениями градиента
vyb <- t(replicate(Nrand, sample(gr, replace=FALSE)))
#  Получение распределения коэффициента a1 для коллекции моделей 1-го порядка на данных
#  с разрушенными статистическими связями
Ra11 <- sapply(1:Nrand, function (i) {
       coef (glm(formula = part.species ~ poly(vyb[i,], 1), family = "binomial"))[2]})
#  Вычисление p-значения для коэффициента a1 модели 1-го порядка
p_a11 <- (sum(abs(Ra11)- abs(coef(glm.1)[2]) >= 0)+1) / (Nrand + 1)
#  Аналогичная проверка статистической значимости коэффициента a2 модели 2-го порядка
Ra22 <- sapply(1:Nrand, function (i) {
       coef (glm(formula = part.species ~ poly(vyb[i,], 2), family = "binomial"))[3]})
p_a22 <- (sum(abs(Ra22)- abs(coef(glm.2)[3]) >= 0)+1) / (Nrand + 1)
```

4. МНОГОМЕРНЫЕ МОДЕЛИ ДИСПЕРСИОННОГО И РЕГРЕССИОННОГО АНАЛИЗА

4.1. Основные модели ANOVA, их ограничения и особенности реализации

В предыдущей главе рассматривались методы статистической оценки совместной изменчивости двух переменных, в том числе, модель одномерной регрессии, при которой вариация значений наблюдаемой случайной величины объясняется влиянием одного независимого значимого фактора. Однако мир по своей природе сложен и многомерен, а ситуации, когда некоторое явление полностью описывается одной переменной, чрезвычайно редки. *Многомерный анализ* (Кендалл, Стьюарт, 1976; Seber, 2004) представляет собой часть статистики, которая выполняет обработку и интерпретацию результатов, полученных на основе наблюдений одновременно нескольких взаимосвязанных случайных переменных, каждая из которых представляется одинаково важной, по крайней мере, первоначально. Например, анализируя продуктивность фитоценоза, нам необходимо рассмотреть множество показателей, связанных с генетической структурой растений, типом почвы, влажностью, освещенностью, температурой и т.д. И здесь информативными могут оказаться как главные влияющие факторы, так и парные эффекты взаимодействия между отдельными факторами, а, возможно, и связи более высших порядков.

У многих из стандартных одномерных методов, таких как однофакторный дисперсионный анализ или простая регрессия, есть многомерные обобщения, которые будут рассмотрены ниже. Помимо этого, есть ряд методов, у которых нет одномерной аналогии, и они будут рассмотрены в последующих главах.

Планы многофакторного эксперимента являются развитием однофакторного ANOVA, когда каждый уровень варьирования одного фактора может комбинироваться со всеми возможными уровнями остальных факторов, причем каждая из этих комбинаций может повторяться. Например, мы можем поставить эксперимент, в котором исследуются влияние температуры ("высокая" и "низкая") и различных видов удобрений ("добавлено" и "не добавлено") на темп роста рассады, выращивая растения при различной температуре и комбинаций подкормки. И если однофакторный дисперсионный анализ ставит своей основной целью проверить статистическую значимость эффекта группировки, то многофакторный анализ вариаций ANOVA смещает акцент в сторону оценки степени влияния отдельных факторов или взаимодействий между ними. В этом смысле он близок множественной регрессии с ее аппаратом селекции наиболее информативного комплекса объясняющих переменных. Поэтому единой концептуальной основой, как многофакторного дисперсионного анализа, так и множественной регрессии является линейная модель 1-го порядка.

Простая линейная модель предполагает, что изменчивость совокупности наблюдаемых объектов может быть объяснена влиянием набора m независимых переменных (x_1, x_2..., x_m) и общим источником стохастических флуктуаций **ε**, обуславливающих нескоррелированную ошибку моделируемой случайной величины: $Y = a_0 + \sum_{i=1}^{m} a_i x_i + \varepsilon$, где a_i, (i = 1,2 …, m) – неизвестные параметры, имеющие смысл масштабирующих коэффициентов при x_i. Как и в одномерном случае, формулируются следующие основные ограничения простой линейной модели:

- значения "ошибки" **ε** представляют собой случайные величины, независимые в совокупности и имеющие одинаковое нормальное распределение $\varepsilon \sim N(0, \sigma^2)$ с нулевым математическим ожиданием и дисперсией $\sigma^2 > 0$;
- наблюдаемые значения случайных величин $Y_1, \ldots, Y_n$ независимы в совокупности и имеют распределения, отличающиеся сдвигом;

- исследуемые факторы влияют независимо и одинаково во всей области определения модели, а их градации сопоставимы и могут быть объединены.

В рамках линейной модели принципиальное отличие между дисперсионным и регрессионным анализом заключается лишь в том, что в первом случае независимые переменные представлены в номинальных шкалах с конечным числом фиксированных категорий. Однако в более сложных ситуациях разница в методологии и особенностях практического применения становится ощутимой.

Под названием дисперсионный анализ фигурирует большая группа методов, которые объединяет один принцип: разложение изменчивости изучаемого показателя на компоненты, объясняемые влиянием независимых внешних факторов и/или их взаимодействий, и случайную ошибку (Закс, 1976; Монтгомери, 1980). Разумеется, при этом не выполняется исчерпывающий перебор всех возможных вариантов построения факторной модели, а только их некоторого подмножества, которое рассматривается как случайное. Многофакторные схемы бывают компактными (состоящими из небольшого числа анализируемых группировок), но по ним нельзя сделать всех выводов, которые возможны и предусмотрены в расширенных схемах. Однако при увеличении числа факторов планы эксперимента становятся громоздкими и не всегда могут удовлетворять требованиям рандомизации (Шитиков и др., 2008).

Является принципиально важным выделение *фиксированных* и *случайных* факторов. Фактор считается фиксированными, если по плану эксперимента набор его уровней находится под контролем исследователя и может быть воспроизведен при реализации повторностей. Если группы измерений выбираются случайно из большого (бесконечного) числа подпопуляций, то фактор считается случайным, а его уровни для каждой повторности формируются из некоторой совокупности возможных градаций.

Линейная модель для двух факторов A и B с повторностями имеет вид:

$$y_{ijk} = \mu + \alpha_i + \beta_j + (\alpha\beta)_{ij} + \varepsilon_{ijk}, \qquad (4.1)$$

где, μ – общее среднее, α_i, β_j – эффекты факторов A и B на $i = 1, \ldots, I$ и $j = 1, \ldots, J$ уровнях воздействия соответственно, y_{ijk} - значение переменной Y, полученное при k-м повторении эксперимента в ячейке ij, $k = 1, \ldots, K$. В таблице полного сбалансированного плана общее число ячеек равно $I \cdot J$, а при выполнении K повторностей эксперимента размерность вектора наблюдений y_{ijk} равна $I \cdot J \cdot K$. Величина $(\alpha\beta)_{ij}$ называется взаимодействием факторов и учитывает эффект комбинации i-го уровня фактора A и j-го уровня фактора B (если он не выражается суммой $\mu + \alpha_i + \beta_j$). Остатки (или случайные невязки) ε_{ijk} предполагаются независимыми и нормально распределенными $N(0, \sigma)$.

Аналогичная линейная модель для трех факторов A, B, C записывается как:

$$y_{ijkl} = \mu + \alpha_i + \beta_j + \gamma_k + (\alpha\beta)_{ij} + (\alpha\gamma)_{ik} + (\beta\gamma)_{jk} + (\alpha\beta\gamma)_{ijk} + \varepsilon_{ijkl},$$

где γ – эффект дополнительного фактора C. Отметим, что, хотя линейные модели для *фиксированных* и *случайных* факторов в целом идентичны, интерпретация ее членов и вид нулевой гипотезы существенно отличаются.

1. Модель с *фиксированными эффектами*. Значение объясняемой переменной Y определяется генеральной средней μ, *дифференциальными* эффектами воздействия индивидуальных факторов, а также комбинаторными эффектами их парных, тройных (и остальных до m) взаимодействий. Например, оценками параметров двухфакторной модели (4.1) являются:

$$\hat{\mu} = \bar{y} \text{ – общее среднее;} \quad \hat{\alpha}_i = \bar{y}_{i\bullet} - \bar{y}; \quad \hat{\beta}_j = \bar{y}_{\bullet j} - \bar{y}; \quad (\alpha\beta)_{ij} = \bar{y}_{ij\bullet} - \bar{y}_{i\bullet} - \bar{y}_{\bullet j} + \bar{x},$$

где точка $\bullet$ означает объединение ячеек плана с данным индексом. При этом для каждого эффекта накладываются ограничения нормировки по всем его уровням

$$\sum_{i=1}^{I} \alpha_i = 0; \sum_{j=1}^{J} \beta_j = 0; \ \sum_{i=1}^{I} (\alpha\beta)_{ij} = 0; \ \sum_{j=1}^{J} (\alpha\beta)_{ij} = 0.$$

Здесь проверяется нулевая гипотеза о том, что соответствующие эффекты не вносят никакого вклада (параметр равен нулю для всех уровней фактора):

H_0(A): $\alpha_1 = \alpha_2 = \ldots = \alpha_I = 0$; H_0(B): $\beta_1 = \beta_2 = \ldots = \beta_J = 0$; H_0(AB): $(\alpha\beta)_{ij} = 0$.

Это эквивалентно другой нулевой гипотезе о равенстве групповых средних:

H_0(A): $\mu_1 = \mu_2 = \ldots = \mu_I$, H_0(B): $\mu_1 = \mu_2 = \ldots = \mu_J = 0$; H_0(AB): $\mu_{ij} = \mu_i + \mu_j - \mu$.

Проверка гипотез может быть осуществлена с использованием дисперсионного *F*-отношения Фишера, в знаменателе которого записывается средний квадрат отклонений для остатков ε.

Классическим примером модели с фиксированными эффектами является эксперимент, в котором одна интактная группа служит контролем, а еще несколько других групп – различными вариантами опыта с заданными уровнями воздействия.

2. Модель *со случайными эффектами*. Используется в случаях, когда исследователя интересует не значимость вкладов уровней факторов, а оценка и сравнение компонентов дисперсии распределения включенных в модель дифференциальных эффектов. Это позволяет обобщить наши заключения о природе факторов и значимости их влияния вне связи с конкретными уровнями, заданными в плане эксперимента. Включение в модель случайного эффекта позволяет выделить из независимых остатков ε с дисперсией σ долю вариации, "в среднем" объясняемую влиянием этого фактора.

В рамках линейной модели (4.1) случайные эффекты α_i, β_j и $(\alpha\beta)_{ij}$ независимы в совокупности и имеют нормальное распределение с нулевым средним и дисперсиями σ_α, σ_β, $\sigma_{(\alpha\beta)}$ соответственно. Проверяются нулевые гипотезы

H_0(A): $\sigma_\alpha = 0$; H_0(B): $\sigma_\beta = 0$; H_0(AB): $\sigma_{(\alpha\beta)} = 0$,

т.е. дисперсия случайного фактора на всех его уровнях равна нулю и не вносит дополнительно вклада в изменчивость наблюдаемой случайной величины.

3. *Модель со смешанными эффектами* включает как фиксированные, так и случайные факторы. Интерпретация членов модели и формулировка нулевых гипотез осуществляется отдельно по спецификациям представленных моделей 1 и 2.

Включение случайных факторов в модели с фиксированными эффектами часто существенно повышает общность заключений о главном результате исследования при распространении его на другие различные пространственные, временные или биологические уровни организации. Например, сделав вывод о влиянии одного фактора (например, добавка или отсутствие удобрения), корректно было бы выполнить анализ о статистической значимости случайных факторов (таких как все возможные места размещения пробных площадок, периоды эксперимента, варианты видов растений и т.д.) по всем возможным их уровням. Эта модель широко используется в биологических исследованиях (Zuur et al., 2009), поскольку многие варианты дисперсионного анализа с повторными измерениями (repeated measures ANOVA) являются смешанными моделями, где случайный фактор – "индивидуальная особь", либо "проба из смеси видов".

Однако неоднозначной и спорной проблемой является выбор знаменателей для *F*-отношений, которые используются для проверки гипотез о статистической значимости случайных факторов в рамках смешанных моделей. Ограничивающий (constrained, restricted) подход полагает, что при вычислении *F*-отношения для случайного фактора, оценивающего значимость прироста дополнительной дисперсии, общий эффект взаимодействия факторов принимается равным нулю. Не ограничивающий (unconstrained) метод выполняет расчет вкладов взаимодействия для каждого уровня случайного фактора с точки зрения их независимости.

Дополнительные трудности возникают при включении в факторную модель двух или более случайных факторов. Уравнения, разрешающие эту проблему, искусственно конструируют выражения для приблизительных оценок знаменателей *F*-отношений (quasi F-ratios). Эти и некоторые другие проблемы не решаются точно в рамках классического подхода и анализируются через регрессионную технику обобщенных линейных (linear

mixed-effects model − LMEM) и нелинейных (NMEM) моделей со смешанными эффектами (Pinheiro, Bates, 2000; Demidenko, 2004; Wood, 2006).

Существуют и иные особенности различных вариантов дисперсионного анализа. По ограничениям, вводимым на взаимодействие между факторами, различают три возможные схемы построения многофакторных моделей:

- перекрёстная схема – классический вариант с полной комбинаторикой взаимодействия факторов (factorial ANOVA);
- иерархическая схема (nested ANOVA), когда выстроена соподчиненность отдельных факторов (например: "регион" → "река" → "станция наблюдения" → "гидробиологическая проба") и, соответственно, градации одного фактора внутри другого не идентичны и поэтому не могут быть объединены в комбинации;
- перекрестно-иерархическая схема (cross-nested ANOVA) – сложные варианты анализа с перекрёстными и вложенными иерархическими эффектами.

Наконец, методологию анализа в значительной мере определяет наличие повторностей и характер распределения общего количества наблюдений по ячейкам, в связи с чем выделяют следующие типы дисперсионного комплекса:

- комплексы с единственным наблюдением в ячейке, когда расчёт статистической ошибки невозможен и в качестве таковой используется взаимодействие факторов;
- равномерные или сбалансированные комплексы, когда в каждой ячейке содержится одинаковое количество наблюдений (наиболее желательный для анализа вариант);
- пропорциональные комплексы, когда соотношение числа наблюдений на всех уровнях фактора одинаковое (например, во всех группах число особей мужского пола ровно в 2 раза больше чем женского);
- неравномерные или несбалансированные комплексы с разным числом наблюдений в ячейках;
- комплексы с пропущенными значениями, когда есть ячейки без наблюдений, которые могут появиться вследствие потери части экспериментального материала или в специальных экспериментальных планах с ограничениями на рандомизацию.

Последние две категории дисперсионных планов также не решаются точно в рамках классического подхода. Модели LMEM-NMEM со смешанными эффектами существенно расширяют возможности общей линейной модели, в том числе:

- разрешается использовать модели с неполными повторными измерениями, когда число наблюдений для разных объектов различно;
- становится возможным анализ коррелированных данных и данных с непостоянной дисперсией;
- можно моделировать не только средние значения, но также дисперсии и ковариации.

Для простой схемы группировки по уровням обобщенную линейную модель LMEM со смешанными эффектами удобно представить в матричной форме:

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{Z}\,\mathbf{b} + \boldsymbol{\varepsilon}, \qquad \mathbf{b} \sim N(\mathbf{0}, \boldsymbol{\psi}_{\boldsymbol{\theta}}), \quad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \Lambda\sigma^2),$$

где $\mathbf{y}$ – N-мерный вектор зависимой переменной, N – число ячеек дисперсионного комплекса, $\mathbf{X}$ и $\mathbf{Z}$ – матрицы, описывающие группировку фиксированных и случайных факторов по уровням в соответствии с планом эксперимента, $\boldsymbol{\beta}$ – вектор фиксированных эффектов (параметров модели). Распределение вектора случайных эффектов $\mathbf{b}$ описывается вектором нулевых средних $\mathbf{0}$ и положительно определенной ковариационной матрицей $\boldsymbol{\psi}_{\boldsymbol{\theta}}$, зависящей от вектора параметров $\boldsymbol{\theta}$, оценка которых являются основной целью статистического вывода о природе случайных эффектов. Наконец, Λ – положительно определенная матрица простой структуры, которая обычно используется при моделировании автокорреляции остатков $\boldsymbol{\varepsilon}$, но часто просто единичная матрица.

Функция правдоподобия L для модели случайных эффектов рассматривается как плотность p распределения вероятности отклика $\mathbf{y}$, определенная параметрами $\boldsymbol{\theta}$, $\boldsymbol{\beta}$, σ :

$$L(\boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2 \,|\mathbf{y}) = p\,(\mathbf{y} \mid \boldsymbol{\beta}, \boldsymbol{\theta}, \sigma^2);$$

максимум L доставляет нам оптимальные оценки $\hat{\theta}, \hat{\beta}, \hat{\sigma}$.

Классические процедуры метода максимального правдоподобия имеют тенденцию занижать оценки параметров. Поэтому многие специалисты предпочитают использовать метод максимального правдоподобия с дополнительными ограничениями на значения параметров (restricted maximum likelihood – REML). Подход REML оценивает качество подгонки параметров модели не с позиций объединения правдоподобий ($\boldsymbol{\theta}$, σ) и $\boldsymbol{\beta}$, а как среднее правдоподобие для всех возможных значений $\boldsymbol{\beta}$, что в рамках байесовского подхода соответствует предположению о локальной однородности распределения для случайных эффектов. REML-критерий, соответствующий среднему правдоподобию, в форме, наиболее удобной для вычислений, может быть представлен как $L_{\mathrm{R}}(\theta, \sigma^2 \mid \mathbf{y}) = \int L(\beta, \theta, \sigma^2 \mid \mathbf{y})\, d\beta$. Подробное изложение теоретических аспектов подбора и интерпретации смешанных моделей с позиций принципов максимального правдоподобия читатель может найти в монографиях (Pinheiro, Bates, 2000; Wood, 2006) или в описании к пакету *mgcv* для статистической среды R, где представлен широкий набор функций для выполнения практических расчетов.

Возможные аргументы использования рандомизационных тестов для многомерных данных – те же самые, что и для одномерных приложений:

- хорошо спланированный эксперимент предполагает случайное распределение экспериментальных единиц по отношению к различным изучаемым факторам;
- если проверяемая нулевая гипотеза верна, то случайные выборки взяты из распределений, которые идентичны между собой;
- нулевая гипотеза подразумевает, что механизм порождения данных имеет равную вероятность возникновения любых их комбинаций, однако, поскольку иерархические и смешанные схемы почти всегда связаны с группировкой, необходимо использовать алгоритмы стратифицированной рандомизации.

Основное различие между одномерными и многомерными тестами в том, что в последнем случае размерность каждой рандомизируемой переменной обычно равна двум или более.

4.2. Выбор модели дисперсионного анализа с фиксированными факторами

Дисперсионный анализ с фиксированными факторами и перекрестной схемой их взаимодействия выполняет разложение общей изменчивости зависимой переменной на составляющие, обусловленные каждым из основных эффектов и всеми возможными комбинациями их взаимодействия.

Рассмотрим принципы трехфакторного дисперсионного анализа на следующем примере [П1]. На 6 станциях наблюдения, расположенных в разных точках акватории Куйбышевского водохранилища, в течение ряда десятилетий брались пробы зоопланктона и оценивалась его средняя биомасса (г/м^3 воды). Измерения выполнялись ежемесячно 6 раз в течение вегетационного периода с мая по октябрь. Динамика средних значений биомассы по месяцам и за различные многолетние периоды представлены на рис. 4.1. Например, можно проследить последовательное уменьшение фауны гидробионтов с 1958 по 1984 гг. вследствие обеднения кормовой базы и влияния антропогенных факторов. Пространственная изменчивость обилия зоопланктона также имеет очевидный характер и будет проиллюстрирована нами далее в главе 7.

С учетом структуры выборочных данных, исходная таблица для проведения дисперсионного анализа состоит из 6×6×3 = 108 ячеек и имеет три входа, соответствующих изучаемым факторам: географическая изменчивость STAN (т.е. список из 6 станций наблюдения), сезонная цикличность MONTH (уровни, соответствующие 6 номерам месяцев в году) и многолетний тренд YEAR, рассматриваемый в контексте трех характерных периодов в истории водохранилища (см. рис. 4.1б). В каждой ячейке располагалось по 6 повторностей измерений за разные годы каждого периода.

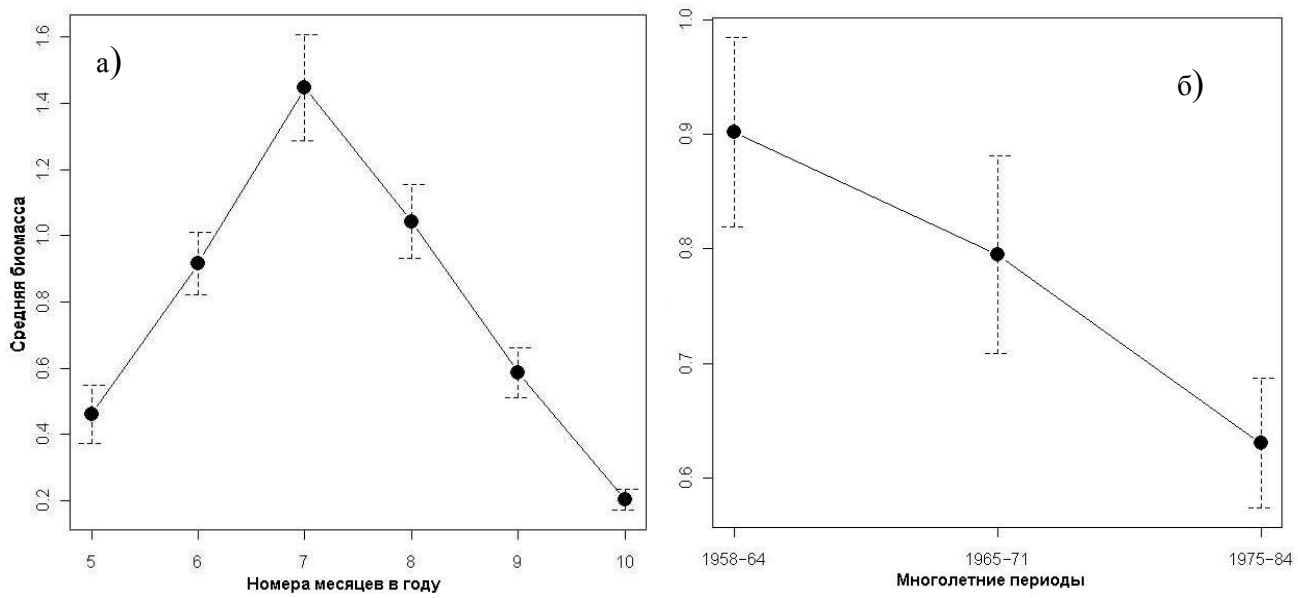


Рис. 4.1. Сезонная а) и многолетняя б) изменчивость биомассы зоопланктона (показаны средние значения и их стандартные ошибки для каждого уровня фактора)

Для формирования полной факторной модели воспользуемся способом, выполняющим перекрестное разложения суммы квадратов, который выделяет главные факторы, все эффекты их парных взаимодействий и эффект совместного действия всех трех факторов:

$$Y = \mu + \text{STAN} + \text{MONTH} + \text{YEAR} + \text{STAN:MONTH} + \text{MONTH:YEAR} + \text{STAN:YEAR} + \text{STAN:MONTH:YEAR} + \varepsilon,$$

где символ ':' связывает факторы, для которых рассчитывается эффект взаимодействия. В результате с использованием программы RTAnova из пакета RT (Б.Манли) получаем следующую стандартную таблицу дисперсионного анализа:

Факторы и их взаимодействия	Сумма квадратов	Степени свободы	*F*-критерий	Оценка *p*-значения Pr(>*F*)	
				параметрическая	рандомизацией
Month	108.74	5	21.87	**< 0.00001**	**0.0002**
Stan	54.48	5	10.95	**< 0.00001**	**0.0002**
Year	8.1	2	4.07	**0.017**	**0.0112**
Month:Stan	14.36	25	0.577	0.951	0.8910
Month:Year	37.76	10	3.79	**0.00005**	**0.0002**
Stan:Year	21.2	10	2.13	**0.0206**	**0.0324**
Month:Stan:Year	40.78	50	0.82	0.805	0.922
Остатки ε (*RSS*)	536.85	540			

Оценки значимости эффектов, полученные после 5000 итераций случайной перестановки значений зависимой переменной относительно комбинаций факторов, соответствуют доле случаев, когда значения *F*-статистик для рандомизированного набора данных оказывалась больше, чем для исходной факторной модели.

Аналогичные результаты могут быть получены с использованием функций aov(), lm() или anova() статистической среды R, которые дополнительно выполняют расчет некоторых важных статистических критериев, а также позволяют осуществить селекцию информативного комплекса факторов. Поскольку модели многофакторного дисперсионного анализа и множественной регрессии основаны на единой концепции общей линейной модели, для их анализа используются одни и те же статистики, которые для нашего примера принимают следующие значения:

- стандартное отклонение для остатков $s_e = (RSS/n)^{0.5} = (536.85/540)^{0.5} = 0.997$;
- коэффициент детерминации $R^2 = 0.347$;

- приведенный коэффициент детерминации $\hat{R}^2 = 1 - \frac{n-1}{n-m}(1-R^2) = 0.217$;
- дисперсионное отношение Фишера $F_{общ} = 2.68$ при 107 и 540 степенях свободы;
- статистическая значимость факторной модели в целом $p < 0.00001$.

Приведенные результаты показывают, что некоторые эффекты взаимодействий факторов, полученные в нашем примере, являются статистически незначимыми. Рекомендуется следующая процедура селекции наилучшей (точнее, "минимальной по количеству используемых членов, но еще адекватной") модели:

1. Формируется полная факториальная модель.
2. Результат исследуется на статистическую значимость отдельных компонентов модели.
3. Формируется новая модель, в которой исключается составляющая с наибольшим *p*-значением (т.е. отбрасывается самый статистически незначимый компонент).
4. Эти две модели сравниваются между собой, и сохраняется новая, более простая модель, если она не вызывает статистически значимого увеличения необъясняемых остатков.
5. Шаги 2-4 последовательно повторяются, пока дальнейшее упрощение модели не станет приводить к излишне грубым оценкам.

Эта процедура пошагового исключения компонентов модели приводит к следующим изменениям ее основных характеристик:

Шаг	Исключаемый компонент	Остатки *RSS*	$R^2_{прив.}$	$F_{общ.}$	$F_{срав.}$	$P_{срав}$
1	Month:Stan:Year	577.63	0.229	4.38	0.82	0.8058
2	Month:Stan	591.99	0.242	7.47	0.586	0.946
3	Stan:Year	613.19	0.228	9.68	2.2	0.016

Здесь $F_{срав}$ – дисперсионное отношение Фишера, с помощью которого оценивается статистическая значимость ($p_{срав}$) превышения остатков *RSS* упрощенной модели по сравнению с остатками модели на предыдущем шаге. В нашем случае модель 3 оказалась существенно менее адекватной, чем модель 2, которая и может быть использована в дальнейшем для содержательной интерпретации:

$$Y = \mu + \text{STAN} + \text{MONTH} + \text{YEAR} + \text{MONTH:YEAR} + \text{STAN:YEAR} + \varepsilon.$$

Автоматический поиск наилучшей модели, доставляющий минимум AIC-критерию и выполненный методом исключений с использованием процедуры boot.stepAIC(), дал следующие результаты, которые не отличаются от полученных выше процедурой селекции "ручным способом":

Шаг	Исключаемый компонент	Остатки *RSS*	Степени свободы	AIC-критерий
0	– (Полная модель)	536.85	540	94.1
1	Month:Stan:Year	577.63	590	41.5
2	Month:Stan	591.99	615	7.42
3	Stan:Year	613.19	625	10.22

Модель, полученная на шаге 3, привела к увеличению AIC-критерия.

С использованием бутстрепа можно существенно расширить состав показателей для анализа значимости компонентов модели. Для этого достаточно сформировать *B* ($B > 1000$) новых наборов данных на основании случайного выбора с возвращением из строк исходной таблицы и для каждого такого псевдонабора осуществить расчет факторной модели. Информативность и устойчивость отдельных коэффициентов модели можно оценить, например, по числу случаев из *B*, при которых: (а) значимость переменной (или эффекта) превысила заданный порог, или (б) коэффициент изменил свой знак.

В ходе итераций бутстрепа могут быть уточнены и такие традиционные критерии, используемые для отбора компонентов модели, как остаточная сумма квадратов *RSS,* приведенный коэффициент детерминации R^2 или информационный критерий Акаике $AIC = -2n \cdot \ln(RSS/n) + 2k$. Возможные смещения коэффициентов финальной модели и их стандартные ошибки удобно оценивать с использованием функции boot(…), реализующей стандартную бутстреп-процедуру для произвольного вектора статистик.

К разделу 4.2:

```
#  Загрузка данных из файла Excel
#  BSuzA <- sqlQuery(channel = 1, select * from [Биомасса зоопланктона$])
#   Или загрузить их из сохраненного двоичного файла
load("BSuzA.RData")
#  Определение факторов
BSuzA$Month <- as.factor(BSuzA$Month)
BSuzA$Stan <- as.factor(BSuzA$Stan)
BSuzA$Year <- factor(BSuzA$Year, levels=c('1958_64','1965_71','1975_84'))
#  Построение трехфакторной модели II типа
Mod1 <- aov(Level ~ Month*Stan*Year, data=BSuzA) ; Anova(Mod1)
#  Последовательное исключение наименее значимых компонентов
Mod2 <-update(Mod1,~.-Month:Stan:Year) ; summary(Mod2) ; anova(Mod1,Mod2)
Mod3 <-update(Mod2,~.-Month:Stan) ; summary(Mod3)  ; anova(Mod2,Mod3)
Mod4 <-update(Mod3,~.-Stan:Year) ; summary(Mod4)  ;  anova(Mod3,Mod4)
# Выполнение автоматической шаговой процедуры
library(bootStepAIC)
boot.stepAIC(Mod1,data=BSuzA)
# Характеристики финальной модели
summary(lm(Level ~ Month + Stan + Year + Month:Year + Stan:Year, data=BSuzA))
# Бутстреп-анализ факторных коэффициентов
library(boot)
bootB <- function(data, indices){data <- data[indices,]
       mod <- lm(Level ~ Month + Stan + Year + Month:Year + Stan:Year, data=data)
       coefficients(mod) }   #  функция, возвращающая вектор коэффициентов
BSuzA.boot <- boot(BSuzA, bootB, 1000) ; BSuzA.boot
```

4.3. Модель со смешанными эффектами и проблема "мнимых повторностей"

Отдельные таксоны донных сообществ, в частности семейство личинок комаров-звонцов Chironomidae, обладают высокой биоиндикационной способностью и широко используются для биотического контроля качества воды (Зинченко, 2011). Поставим своей целью оценить статистическую значимость зависимости популяционной плотности организмов подсемейства Ortocladeiinae от уровня загрязнения водотоков.

Для проверки этой гипотезы будем использовать результаты взятия 557 гидробиологических проб из 33 малых и средних рек в лесостепной части Волжского бассейна [пример П2]. По данным этих наблюдений особи Ortocladeiinae встретились в 285 пробах из общего числа, а выборочное распределение популяционной плотности Y (экз/м^2) характеризуется сильной левосторонней асимметрией, обусловленной обширным нулевым "хвостом". Рассмотрим в составе модели четыре потенциально значимых фактора, каждый из которых имеет по три условных градации:

- оценки качества воды (Water) по гидрохимическим и микробиологическим показателям в соответствие с требованиями ГОСТ 17.1.3.07–82, подробно обсуждаемые в статье (Шитиков, Зинченко, 2005): "1. Относительно чистые" (менее 3.4); "2. Умеренно грязные" (от 3.4 до 4.4); "3. Сильно загрязненные" (свыше 4.4);
- тип грунта (GroundType): "1. Песчано-галечный"; "2. Глинисто-илистый"; "3. Темный ил";
- категория участка реки (RiverType): "1. Ручьи и узкие реки" (ширина менее 10 м); "2. Малые реки" (от 10 до 30 м); "3. Средние реки" (свыше 30 м);
- период взятия проб (Season): "1. Май-июнь"; "2. Июль"; "3. Август-Сентябрь".

Баланс частот наблюдений по категориям факторов сильно деформирован, а для трех ячеек имеются пропуски наблюдений, т.к. в обследованном регионе отсутствует хотя бы один участок реки шириной более 30 м с относительно чистым качеством воды. Заполнение ячеек факторного плана повторностями проб представлено в табл. 4.1.

Таблица 4.1. Изменение популяционной плотности Ortocladeiinae $\ln(Y+1)$, Y – численность экз/м^2 , по данным гидробиологической съемки, произведенной на 33 малых и средних реках Среднего Поволжья

Качество вод	Тип грунта	Категория участка реки	Общее (n) число проб	Проб (n_{or}) с наличием ортокладеин	Плотность популяции	
					Средняя $\ln(Y+1)$	Приведенная $\ln(Y+1)n_{or}/n$
1. Относительно чистые	1.Песчано-галечный	1. Ручьи…	62	52	4.67	3.92
		2. Малые…	4	2	2.72	1.36
		3. Средние…	-	-	-	-
	2. Глинисто-илистый	1. Ручьи…	36	28	4.14	3.22
		2. Малые…	8	7	4.14	3.62
		3. Средние…	-	-	-	-
	3. Темный ил	1. Ручьи…	25	23	4.84	4.45
		2. Малые…	10	6	3.58	2.15
		3. Средние…	-	-	-	-
2. Умеренно грязные	1.Песчано-галечный	1. Ручьи…	33	21	3.39	2.16
		2. Малые…	33	12	1.64	0.598
		3. Средние…	12	3	1.36	0.341
	2. Глинисто-илистый	1. Ручьи…	24	14	2.67	1.56
		2. Малые…	42	26	2.95	1.82
		3. Средние…	13	4	1.27	0.391
	3. Темный ил	1. Ручьи…	19	11	2.69	1.56
		2. Малые…	50	11	1.00	0.221
		3. Средние…	14	5	1.51	0.540
3. Сильно загрязненные	1.Песчано-галечный	1. Ручьи…	2	2	4.76	4.76
		2. Малые…	13	7	2.09	1.13
		3. Средние…	39	14	1.602	0.576
	2. Глинисто-илистый	1. Ручьи…	1	0	0	0
		2. Малые…	1	0	0	0
		3. Средние…	49	21	1.80	0.771
	3. Темный ил	1. Ручьи…	3	2	2.55	1.699
		2. Малые…	8	3	1.68	0.629
		3. Средние…	56	11	0.986	0.194

В практике гидробиологических исследований часто используются различные индексы, являющиеся некоторыми функциональными преобразованиями от наблюдаемой численности организмов, которые служат для обобщения результатов и компенсации эффекта отклонений от нормального закона распределения. Рассмотрим модели дисперсионного анализа на основе двух многофакторных планов:

- сбалансированного трехфакторного плана без учета сезонности с числом ячеек $3\times3\times3 = 27$, в каждую из которых помещена средняя прологарифмированная приведенная численность ортокладеин $\ln(Y+1)\cdot n_{or}/n$, т.е. с дополнительным учетом доли проб, в которых встретились организмы этого подсемейства при каждой комбинации анализируемых факторов (в пропущенные ячейки внесено значение 0);
- несбалансированного плана с пропусками, который включает все четырех фактора и учитывает все множество из 557 наблюдений, причем повторностями являются реально найденные численности организмов в пробах $\ln(Y+1)$.

Для перекрестной модели 1 с фиксированными эффектами на основе приведенных численностей единственным статистически значимым фактором явилась категория участка реки, откуда выполнялись гидробиологические пробы – см. табл. 4.2.

Таблица 4.2. Результаты дисперсионного анализа изменчивости численности ортокладеин с использованием различных моделей с фиксированными (F) и случайными (RE) эффектами

Эф-фект	Факторы и их взаимодействия	Сумма квадратов	Степени свободы	Средние квадраты	Тестовая-статистика	Оценка *p* Pr(>*T*)
Модель 1 – фиксированная по приведенной численности и сбалансированному плану						
F	Water	4.46	1	4.46	4.12	0.0545
	GroundType	0.642	1	0.642	0.594	0.449
	Water:GroundType	2.31	1	2.31	2.14	0.157
	RiverType	23.37	1	23.37	21.61	**0.00012**
Остатки ε		23.79	22	1.082		
Модель 2 – смешанная по приведенной численности и сбалансированному плану						
F	Water	4.46	1	4.46	2.74	0.056
	GroundType	0.642	1	0.642	-0.347	0.487
	Water:GroundType	2.31	1	2.31	1.503	0.177
RE	RiverType			1.183		
Остатки ε				1.124		
Модель 3 – фиксированная по численности в пробах и несбалансированному плану						
F	Water	624.3	1	624.3	102.3	< **0.00001**
	GroundType	45.7	1	45.7	7.48	**0.0064**
	Water:GroundType	0.92	1	0.92	0.15	0.7
	RiverType	70.5	1	70.5	11.5	**0.00072**
	Season	85.0	1	85.0	13.9	**0.00035**
Остатки ε		3360.1	551	6.1		
Модель 4 – смешанная по численности в пробах и несбалансированному плану						
F	Water	113.1	1	113.1	15.89	**0.001**
	GroundType	24.8	1	24.8	1.854	**0.042**
	Water:GroundType	4.62	1	4.62	-0.946	0.386
RE	RiverType			0.74		
	Season			0.51		
Остатки ε				2.44		

Примечание: В качестве тестового критерия *T* использовались *F*-отношение (модель 1), χ^2 (модель 3), а для моделей 2 и 4 со смешанными эффектами – LRT-статистика (likelihood ratio test или отношение оценок максимального правдоподобия), *p*-значение которой оценивались по результатам параметрического бутстрепа.

Здесь мы столкнулись с проблемой **"**мнимых повторностей**"** (или псевдорепликации), обсуждаемой С. Хелбертом, перевод статьи которого (Hulbert, 1984) вместе с подробными комментариями представлен в нашем сборнике «Проблемы экологического эксперимента» (2008). Псевдорепликация является одним из наиболее широко цитируемых и недооцененных на практике понятий в статистическом анализе экологических исследований и определяется как «использование дедуктивной статистики для оценки влияния фактора, когда данные эксперимента фактически не имеют повторностей или эти повторности не являются статистически независимыми...». Действительно, в нашем случае мы ставим своей целью выявить влияние качества воды на популяционную плотность ортокладеин, но этот показатель зависит в свою очередь от сезонного периода и гидрологических характеристик участка реки, откуда брались пробы.

Л. Чавес в своей статье с характерным названием «Руководство энтомологу по демистификации псевдорепликации» (Chaves, 2010) предложил использовать обобщенную линейную модель со смешанными эффектами (LMEM) как статистическое решение по анализу данных полевых исследований с псевдоповторностями. В этом случае можно корректно разделить между собой различные источники изменчивости зависимой

переменной и получить правильные выводы из данных, собранных в исследованиях с ограничениями на рандомизацию.

Расчеты в статистической среде R были реализованы нами с использованием функции lmer(…) из пакета lme4, а коды скриптов представлены в дополнении к этому разделу. Отличие полученной модели 2 со смешанными эффектами

$$Y_{\text{пр}} = \mu + \text{Water} + \text{GroundType} + \text{Water:GroundType} + RE(\text{RiverType}) + \varepsilon$$

от аналогичной модели 1 с фиксированными эффектами

$$Y_{\text{пр}} = \mu + \text{Water} + \text{GroundType} + \text{Water:GroundType} + \text{RiverType} + \varepsilon$$

состоит в том, что при объявлении RiverType случайным фактором все обследованное множество рек интерпретируется как случайная выборка из генеральной совокупности водных объектов, а ее вклад представлен в модели значением вариации *RE*(RiverType), которая аккумулирует в себе всю возможную изменчивость категорий водотоков. Иными словами, тип рек перестает быть независимым предиктором модели, но, оставаясь компонентом разложения дисперсии, наряду с остатками ε ограничивает возможность совершения ошибки 1-го рода, то есть отклонить нулевую гипотезу, когда она верна.

Параметры обобщенной линейной модели со смешанными эффектами (LMEM) могут быть рассчитаны с использованием общего принципа максимального правдоподобия, а их статистическая значимость может быть протестирована с использованием параметрических статистик (F, χ^2), что полностью корректно только для сбалансированных планов. Однако метод максимального правдоподобия с ограничениями на значения параметров (REML) в сочетании с параметрическим бутстрепом предоставляет более общий подход, корректный для различных версий дисперсионной модели.

При реализации этого подхода последовательно для каждого фиксированного фактора *f* дисперсионного комплекса выполняются следующие действия. Предварительно по исходной выборке наблюдений строятся две модели – полная $\mathbf{M}_{\text{full}}$ и без одного тестируемого члена $\mathbf{M}_{\text{m1}}$. Для каждой из этих моделей вычисляются оценки максимального правдоподобия и находится логарифм их отношения $\text{LRT}_{\text{obs}} = 2\ln(L_{\text{full}} / L_{\text{m1}}) = 2(\ln L_{\text{full}} - \ln L_{\text{m1}})$. Чтобы оценить статистическую значимость отношения LRT_{obs}, которое соответствует мере снижения качества модели, если из нее исключить один фактор, выполняется следующая процедура параметрического бутстрепа:

- извлекается случайная выборка из распределения, порождаемого моделью $\mathbf{M}_{\text{m1}}$;
- эта выборка используется в качестве зависимой переменной и рассчитываются две нуль-модели с использованием формул моделей $\mathbf{M}_{\text{full}}$ и $\mathbf{M}_{\text{m1}}$;
- вычисляется логарифм отношения оценок максимального правдоподобия $\text{LRT}_{\text{sim}} = 2\ln (L_{\text{full\_sim}} / L_{\text{m1\_sim}})$ этих нуль-моделей (т.е. поскольку обе модели создавались на основе одной и той же выборки, сгенерированной в соответствии с ограничениями модели $\mathbf{M}_{\text{m1}}$, где влияние фактора *f* элиминировано, то их отличия в качестве подгонки данных должны носить случайный характер);
- повторяются вышеперечисленные шаги *B* (т.е. несколько тысяч) раз и строится бутстреп-распределение отношения оценок максимального правдоподобия LRT_{sim};
- если эмпирическое значение LRT_{obs} превышает верхний уровень доверительного интервала LRT_{sim}, то включение фактора в модель статистически значимо.

Важнейшим преимуществом обобщенной линейной модели LMEM является статистически корректный анализ несбалансированных планов с пропущенными значениями. В представленном примере мы получаем возможность выполнить дисперсионный анализ с включением всех четырех факторов по всему множеству из 557 наблюдений. При этом повторностями в каждой ячейке плана являются реально найденные численности организмов в пробах, а не некоторые средние или приведенные индексы, полученные по субъективным формулам. Модель 3 (табл. 4.2) с фиксированными эффектами была построена на основе базового распределения Пуассона (family = poisson), а таблица дисперсионного анализа получена с использованием критерия

χ^2. Все четыре индивидуальных фактора, используемых для группировки, оказались статистически значимыми, что не вполне приводит к корректному выводу по существу нашей задачи.

Построим теперь модель 4 со смешанными эффектами и объявим случайными факторами оба показателя, которые оказывают влияние на независимость повторностей, но не относятся к сути решаемой проблемы: тип рек RiverType и период взятия проб Season. Тогда их средние квадраты (см. табл. 4.2) сосредотачивают в себе пространственно-временную изменчивость комплекса и в некотором смысле обеспечивают статистическую независимость проявления целевых эффектов. Низкие *p*-значения ($p < 0.05$), полученные для фиксированных факторов модели 4, позволяют нам сделать вывод о значимом влиянии на популяционную плотность ортокладеин, в первую очередь, качества воды Water и, в меньшей степени, типа грунта.

К разделу 4.3:

```
#### Линейная модель LM и модель со смешанными эффектами LMEM
# Загрузка данных из текстовых файлов
data1=read.table("Orto_Bal.txt",header=T,sep="\t")  #  средние для сбалансированного плана
# Загрузка по несбалансированному плану  из текстового файла
data2=read.table("Orto_Full.txt",header=T,sep="\t")  #  полные данные численности
### Построение линейных моделей (1) и (3)
model1=lm(Adjus _Orto~Water*GroundType+RiverType,data1) ; anova(model1)
model3=lm(log(Orto+1)~Water*GroundType+RiverType+Season,data2,family = poisson)
summary(model3) ; anova(model3, test="Chi")
### Построение модели со смешанными эффектами (1) и (3)
Nrep=1000            #  Задаем число перевыборок бутстрепа
# Функция оценки p-значения для терма смешанной модели параметрическим бутстрепом
#  mod_full – полная модель; mod_m1 – модель с одним исключенным членом
P_value <- function (mod_full, mod_m1, data) {   lrsl = matrix(0,Nrep)
  #  Корректировка формул модели
  f_full <- as.formula(paste("sim_m1 ~",strsplit(as.character(formula(mod_full)),"~")[3]))
  f_m1 <- as.formula(paste("sim_m1 ~",strsplit(as.character(formula(mod_m1)),"~")[3]))
  #   Извлечение выборки из распределения
   for (i in 1:Nrep)   { sim_m1 = unlist(simulate(mod_m1))
       #   Построение новых моделей на основе имитируемой выборки
     smod_full=lmer(formula = f_full,data) ;       smod_m1=lmer(formula = f_m1,data)
       #  Получение бутстреп-распределения разностей оценок максимального правдоподобия
     lrsl[i]= 2*(logLik(smod_full)-logLik(smod_m1))[1] }
       #  Разность оценок логарифмов максимального правдоподобия для исходных моделей
       thresh <- 2*(logLik(mod_full)-logLik(mod_m1))[1]
       return (mean(lrsl>thresh)) }
# ---------------------------------------------------------------------
### Подбор модели (2), где (1|factor) определяет случайный фактор
Model2=lmer(Adjus_Orto~Water*GroundType+(1|RiverType),data1)
###  Подбор модели без взаимодействия
Model2_1=lmer(Adjus_Orto~Water+GroundType+(1|RiverType),data1)
### Упрощенные модели для оценки значимости Water и GroundType
Model2_2=lmer(Adjus_Orto~Water+(1|RiverType),data1)
Model2_3=lmer(Adjus_Orto~GroundType+(1|RiverType),data1)
#   Оценка p-значений  и LRT-статистик
P_value(model2, model2_1, data1) ; 2*(logLik(model2)-logLik(model2_1))[1]
P_value(model2_1, model2_2, data1) ; 2*(logLik(model2_1)-logLik(model2_2))[1]
P_value(model2_1, model2_3, data1) ; 2*(logLik(model2_1)-logLik(model2_3))[1]
### Построение модели (4) со смешанными эффектами по несбалансированному плану
Model4=lmer(log(Orto+1)~Water*GroundType + (1 | RiverType) + (1 | Season),data2) ;
anova(model0)
#  Построение серии моделей с различными комбинациями включаемых эффектов
model4_1 =lmer(log(Orto+1)~Water + GroundType + (1 | RiverType) + (1 | Season),data2)
model4_2 =lmer(log(Orto+1)~Water + (1 | RiverType)+ (1 | Season) ,data2)
model4_3 =lmer(log(Orto+1)~GroundType + (1 | RiverType) + (1 | Season),data2)
#   Оценка p-значений  и LRT-статистик
```

```
P_value(model4, model4_1, data2) ; 2*(logLik(model4)-logLik(model4_1))[1]
P_value(model4_1, model4_2, data2) ; 2*(logLik(model4_1)-logLik(model4_2))[1]
P_value(model4_1, model4_3, data2) ; 2*(logLik(model4_1)-logLik(model4_3))[1]
```

4.4. Иерархический (гнездовой) дисперсионный анализ

При экологическом мониторинге широко распространена иерархическая схема организации наблюдений, когда, например, совокупность гидробиологических проб, взятых из фиксированного биотопа (site), является одной из серий измерений на некотором целостном участке реки (area), которые потом совместно обобщаются с аналогичными блоками данных по другим участкам водотока. Можно продолжить агрегирование данных и далее, когда мероприятия по мониторингу конкретной реки входят в состав регионального плана обследования обширной водохозяйственной сети и т.д. Каждый из перечисленных объектов иерархии является фактором, оказывающим влияние на изменчивость анализируемого показателя: на рис. 4.2 это отдельные участки **A** в составе реки и станции наблюдения **B** на каждом из участков [пример П2]. Специфика такого плана эксперимента состоит в том, что факторные пространства являются непересекающимися по горизонтали, т.е. отсутствует какая-нибудь связь между уровнями фактора **B**, принадлежащими разным уровням фактора **A** (например, между станциями наблюдения двух разных участков). При этом фактор **A** является главным по отношению к вложенному ("минорному") фактору **B**.

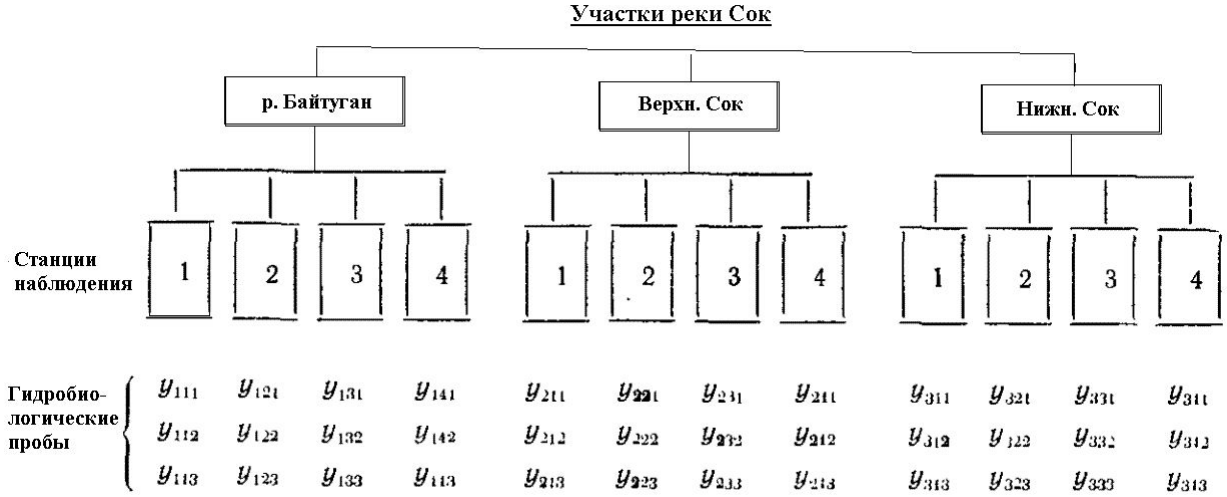


Рис. 4.2. Вид двухступенчатого гнездового плана

При иерархическом плане оцениваются две составляющие изменчивости: (а) различия между отдельными участками, которые могут быть обусловлены, например, воздействием выявляемых факторов окружающей среды и (б) варьирование показателя между повторностями измерений внутри каждого участка в результате, например, пространственно-временной неоднородности. Дисперсионный анализ (nested ANOVA) такого гнездового двухфакторного плана также может быть основан на общей линейной модели, которая для любого случайно взятого измерения записывается теперь как

$$y = \mu + \tau + \beta|\tau + \varepsilon,$$

где μ – математическое ожидание общего среднего; τ – влияние фактора изучаемого воздействия; $\beta|\tau$ – влияние изменчивости наблюдений внутри групп β с одинаковым уровнем воздействия τ; ε – влияние случайных (не учтенных в эксперименте) факторов.

Предполагается, что все факторы τ, $\beta|\tau$ и ε независимы друг от друга, поэтому общую суммарную изменчивость можно разложить на три компоненты:

$$MS_y = MS_\tau + MS_{\beta|\tau} + MS_\varepsilon.$$

Значимость средних квадратов MS, соответствующих этим компонентам, можно проверить по критерию Фишера. Если отношение $F_1 = MS_\tau / MS_{\beta|\tau}$ статистически значимо, то можно говорить о достоверном влиянии главного фактора группировки τ. В противном случае изменчивость анализируемого показателя может быть объяснена влиянием вложенного фактора β|τ (при статистической значимости дисперсионного отношения $F_1 = MS_{\beta|\tau} / MS_\varepsilon$), либо случайными погрешностями измерений ε.

В качестве примера выполним отбор из базы данных 120 гидробиологических проб макрозообентоса по схеме сбалансированного двухступенчатого гнездового плана, т.е. для каждого створа реки (станции) назначим по 10 повторностей наблюдений. Используем дисперсионный анализ для сравнения видового разнообразия макрозообентоса по индексу Шеннона между отдельными участками рек – см. рис. 4.3.

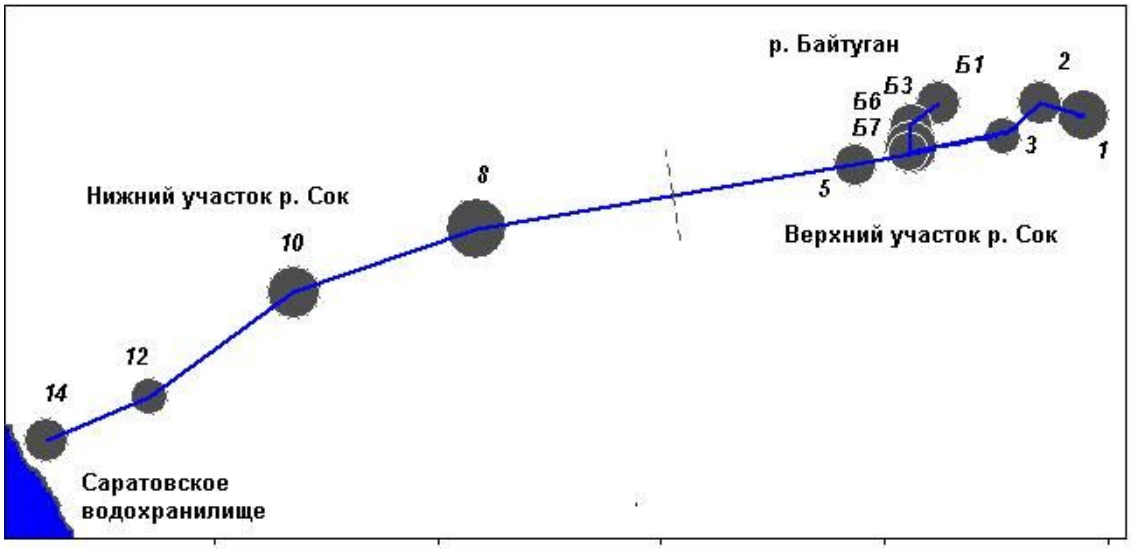


Рис. 4.3. Схема расположения участков и станций наблюдения на р. Сок (диаметр кружков соответствует уровню видового разнообразия макрозообентоса по индексу Шеннона)

Статистический вывод о влиянии фиксированного фактора **A** будет касаться конкретных градаций деления водотока на три участка, которые мы задали априорной группировкой. Случайный фактор **B**{**A**}, вывод о котором распространяется на все возможные его уровни, будет отражать влияние различий между гидробиологическими пробами на станциях внутри каждого участка. Отдельно выделим долю случайной вариации показателя внутри групп фактора **B** (и тем самым внутри групп фактора **A**) – "остаток {error}" в табл. 4.3.

На первом этапе рассчитываем сумму квадратов SS_A отклонений, обусловленную фиксированным фактором **A**, а остальную сумму квадратов раскладываем на две компоненты: $SS_{B(A)}$ и SS_{error}. Далее проверяется нулевая гипотеза H_0:"Нет различий в видовом разнообразии между участками водотока", для чего факториальный средний квадрат MS_A сравнивается с вариансой $MS_{B|A}$, определяющей изменчивость биотопов внутри выделенных участков рек. По имеющимся данным эту гипотезу отклонять нельзя.

Таблица 4.3. Результаты сравнения средних квадратов отклонений, полученных в ходе двухфакторного иерархического анализа

Факторы	Эффект	Степени свободы *df*	Сумма квадратов *SS*	Средние квадраты *MS*	*F*-критерий	*p*-значение	
						параметр.	рандомиз.
A (Участок)	Фиксированный	2	3.16	1.58	0.66	0.539	0.550
B{**A**} Станция	Случайный	9	21.55	2.39	2.053	**0.04**	**0.035**
Остаток (error)		108	125.9	1.16			

На следующем этапе проверяется вторая нулевая гипотеза, которая звучит как H_0:"Нет различий в видовом разнообразии для створов водотока в пределах выделенных участков". Для этого средние квадраты $MS_{B|A}$ вложенного фактора сравниваются с остаточной дисперсией MS_{error}, определяющей внутреннюю видовую изменчивость донного сообщества в повторностях проб внутри каждой станции. Эта гипотеза отвергается на выбранном уровне значимости ($\alpha = 0.05$). Таким образом, влияние изучаемого фактора – гидрологических и географических особенностей – можно считать недоказанным на фоне изменчивости разнообразия зообентоценозов внутри и между отдельными биотопами. Для уточнения научного предположения о существовании продольного градиента реки требуется дополнительный эмпирический материал.

Оценка *p*-значений для *F*-критерий выполнялась двумя способами: параметрическим на основе теоретического *F*-распределения и с использованием рандомизации. Для главного фактора **A** случайная перестановка значений *y* проводилась без ограничений на всем массиве из 120 наблюдений (см. скрипт в кодах R, представленный ниже). Для вложенного фактора **B** выполнялась рандомизация с ограничениями: значения индексов Шеннона перемешивались только в пределах проб, взятых на каждом из трех участков водотока.

К разделу 4.4:

```
# Загружаем данные из xls-файла
library(xlsReadWrite)
# Таблица численностей "виды-пробы"
TTB <- t(read.xls("Manova.xls", sheet = 1, rowNames=TRUE))
# Таблица "привязки" проб к створам и участкам рек
TTB.Site <- read.xls("Manova.xls", sheet = 2, rowNames=TRUE)  ; attach(TTB.Site)
#  Загрузка относительных географических координат станций
TTB.Coor <- read.table("Sok_coordinates.txt",header=F) ; colnames(TTB.Coor)<- c("X","Y")
rownames(TTB.Coor)<- c("Б_01","Б_03","Б_06","Б_07","С_01","С_02","С_03","С_05",
                                                    "С_08","С_10","С_12","С_14")
library(vegan)
Shennon <- diversity(exp(TTB)) #  Расчет индекса Шеннона для каждой гидробиологической пробы
#  Расчет среднего индекса Шеннона для 10 повторностей проб на каждой станции
T <- cbind(TTB.Site, Shennon) ; MShennon <- tapply(T$Shennon, T$Site, FUN=mean)
#  Отрисовка графика расположения станций в географических координатах
#  Рисуются круги, пропорциональные среднему индексу Шеннона на каждой станции
plot(TTB.Coor, asp=1,  pch=21, col="white", bg="grey30", cex=5*MShennon/max(MShennon))
lines(TTB.Coor, lwd=2, col="blue") ; text(TTB.Coor, row.names(TTB.Coor), cex=0.5,
                                                          col="white")
#  Сохраняем созданные объекты для повторного использования
save("TTB", "TTB.Coor", "TTB.Site", file="Sok.RData")
# Иерархический (гнездовой) ANOVA с одним главным и одним вложенным фактором
# (адаптация скрипта D.Borcard, P.Legendre,2007)
River.f = as.factor(River)   # Главный фактор
Site.f = as.factor(Site)     # Вложенный фактор
dat = as.data.frame(cbind(Shennon, River.f, Site.f))
# Проверка сбалансированности плана: одинаковое число повторностей для каждого блока
balance = var(as.vector(table(dat[,2])))
if(balance > 0) stop("План несбалансирован. ")
# Выполняем формирование линейной модели и таблицы дисперсионного анализа
# Иерархия вложенности при записи формулы определяется %in%
anova.res = anova(lm(Shennon ~ River.f + Site.f %in% River.f))
if(nrow(anova.res)<3) cat("Проблемы с данными (отсутствие повторностей ?)",'\n
            ',"Пермутационный тест выполняется только для главного фактора",'\n')
# Пересчитываем параметрическую  F-статистику и ассоциируем вероятность с главным фактором
F.main = anova.res[1,3] / anova.res[2,3]
P.main = pf(F.main, anova.res[1,1], anova.res[2,1], lower.tail=FALSE)
anova.res[1,4] = F.main ; anova.res[1,5] = P.main
# Пермутационный тест
# Функция, возвращающая рандомизированный вектор порядковых номеров объектов
```

```
restrictedPerm <- function(nobs.block, nblock, n, restPerm=TRUE, vec)   {
# restPerm == F: Неограниченная рандомизация.
# restPerm == T: Рандомизация с ограничениями, т.е. наблюдений внутри каждого блока данных.
# Например:  toto0 <- restrictedPerm(6,4,24,FALSE,c(1:24))
       if(restPerm == FALSE) { vec <- sample(vec[1:n],n)
          } else {
          for(j in 1:nblock) {
             i1 <- nobs.block*(j-1)+1 ;  i2 <- nobs.block*j
             vec[i1:i2] <- sample(vec[i1:i2],nobs.block)
             }    }    return(vec) }
#  ------------------------------
nperm = 1000
n = nrow(dat) ; vec = seq(1:n) ; nblock = length(levels(River.f))
nobs.block = nrow(dat)/nblock ; k = nrow(anova.res) - 1
Pperm = c(rep(0,k), NA)
# Для вложенного фактора данные переставляются внутри каждого уровня главного фактора
if(nrow(anova.res)>=3){
    data2 = dat[order(dat[,2]),]  # Сначала данные сортируются по значению главного фактора
              GEn = 1
              for(i in 1:nperm){
       # Рандомизация с ограничениями
                     vecperm = restrictedPerm(nobs.block, nblock, n, restPerm=TRUE, vec)
                     Y.perm = data2[order(vecperm),1]
                     anova.perm = anova(lm(Y.perm ~ River.f + Site.f %in% River.f))
                     if(anova.perm[2,4] >= anova.res[2,4]) GEn = GEn + 1  }
              Pperm[2] = GEn/(nperm+1)
}
# Рандомизационный тест для главного фактора
GEm = 1  ;    for(j in 1:nperm){
       Y.perm = sample(Shennon, n)
       anova.perm = anova(lm(Y.perm ~ River.f + Site.f %in% River.f))
       F.main.perm = anova.perm[1,3] / anova.perm[2,3]
       if(F.main.perm >= F.main) GEm = GEm + 1  }
Pperm[1] = GEm/(nperm+1)
#  Вывод результатов
anova.res  = data.frame(anova.res, Pperm)
colnames(anova.res) = c("Df", "Sum Sq", "Mean Sq", "F value", "Prob(param)", "Prob(perm)")
note = "Иерархический ANOVA, параметрический и рандомизационный тест"
list(anova.type=note, nperm=nperm, anova.table=anova.res)  # Вывод результатов
#  Сохраняем таблицы для использования в других скриптах
save (TTB, TTB.Site, file="Sok.RData")
```

4.5. Модель множественной линейной регрессии

Если рассматривается влияние несколько ($m > 1$) факторов на зависимую переменную Y, то естественным обобщением парной регрессии (см. раздел 3.3) является многомерная регрессионная модель (multiple regression model):

$E(Y|x_1, x_2, \ldots, x_m) = f(x_1, x_2, \ldots, x_m)$, где $f(\ldots)$ – произвольная функция m переменных.

Самой употребляемой и наиболее простой является модель линейной регрессии:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_m X_m + \varepsilon. \qquad (4.2)$$

Каждый коэффициент регрессии β_j, $j = 1, 2, \ldots, m$, численно отражает степень влияния предиктора X_j на условное математическое ожидание $M(Y|x_1, x_2, \ldots, x_m)$ зависимой переменной Y, если мысленно принять, что все остальные объясняющие переменные модели остаются постоянными. Относительно случайных ошибок $\boldsymbol{\varepsilon}$ (или остатков модели) в классическом случае делаются следующие предположения, являющиеся предпосылками теоремы Гаусса-Маркова:

- статистическая независимость и некоррелированность для разных наблюдений $E(\varepsilon_i^2) = \sigma^2$, $E(\varepsilon_i \varepsilon_j) = 0$ при $i \neq j$;

- $E\varepsilon_i = 0$ и совместное нормальное распределение $\varepsilon_i \sim N(0, \sigma^2)$.

Как и в случае парной регрессии, анализ начинается с расчета величин b_j, являющихся выборочными оценками коэффициентов модели β_j, j = 1, 2, …, m. Далее выполняется проверка статистических гипотез, позволяющая сделать вывод о надежности найденных точечных оценок коэффициентов, интерпретируемости найденного эмпирического уравнения регрессии и степени его адекватности результатам наблюдений.

Для проверки гипотезы H_0: $b_j = \beta_j$ используется, как правило, статистика: $t = (b_j - \beta_j)/S_{bj}$, где S_{bj} – ошибка коэффициента регрессии, равная стандартному отклонению случайной величины b_j. Поскольку t при справедливости H_0 имеет распределение Стьюдента, нетрудно найти соответствующее ему p-значение. Наряду с проверкой статистической значимости, выполняется оценка доверительных интервалов β_j.

Качество полученной модели принято оценивать по следующим критериям:

- стандартное отклонение для остатков $S_e = \sqrt{(\sum e_i^2)/(n-m-1)}$, где **e** – вектор отклонений выборочных значений y_i зависимой переменной Y от значений $\hat{y}_i$, получаемых по уравнению регрессии;
- коэффициент детерминации $R^2 = 1 - \sum e_i^2 / \sum (y_i - \bar{y})^2$;
- приведенный коэффициент детерминации $\overline{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-m-1}$, который не столь быстро растет при добавлении новых термов в регрессионную модель;
- F-статистика $F = \frac{R^2}{1-R^2} \cdot \frac{n-m-1}{m}$, p-значение для которой может быть получено с использованием распределения Фишера;
- информационный критерий Акаике $AIC = n \ln(\sum e_i^2 / n) + 2k$ и некоторые его "клоны" – Байесовский информационный критерий (BIC), критерий Шварца и др., представленные в разделе 3.4.

Ниже на примере будут рассмотрены основные пункты протокола статистических процедур, обеспечивающих построение хорошо интерпретируемых моделей:

- анализ мультиколлинеарности (или взаимосвязи) независимых переменных;
- отбор информативных факторов многомерной функции регрессии;
- регрессионная диагностика и оценка статистической значимости параметров;
- анализ степени нелинейной связи между переменными.

Проблема *мультиколлинеарности* приводит к тому, что матрицы, используемые при вычислении коэффициентов, становятся плохо обусловленными, оценки параметров оказываются неустойчивыми и трудно анализировать вклад каждого отдельного фактора в прогнозируемую величину. Степень мультиколлинеарности j-го каждого признака, j = 1, 2, …, m, может быть оценена *фактором роста дисперсии* (Variance Inflation Factor, VIF):

$$\mathrm{VIF}(\beta_j) = 1/(1 - R_j^2),$$

где R_j^2 – коэффициент детерминации для регрессии признака X_j по всем оставшимся (m - 1) независимым переменным. Этот показатель имеет вполне определенный диагностический смысл: текущая дисперсия выборочной оценки коэффициента регрессии β_j в $\mathrm{VIF}(\beta_j)^{0.5}$ раз превышает идеальную оценку, если бы взаимосвязи между переменными не было. Значение VIF_j от 1 до 2 (что соответствует R_j^2 от 0 до 0.5) означает отсутствие проблемы мультиколлинеарности для X_j, т.е. добавление или удаление других независимых переменных не изменяет оценки коэффициента b_j и статистики t_j.

В дельте реки Волги [пример П3] на 159 участках изучался состав травянистого покрова и оценивалась суммарная надземная биомасса (Bmass) растений (тростника, рагозы, пырея, череды, алтея и др.) в г на м^2 площади. По результатам химического анализа водной вытяжки определялся ионный состав почв: хлориды (Cl), сульфаты

(Sulfat), бикарбонаты (Carbon), кальций (Ca) , натрий (Na) и магний (Mg) в мг-экв. Для каждого участка измерялась средняя высота над уровнем межени (H), которая имеет сильную корреляцию со степенью увлажненности грунта.

Рассчитаем предварительно матрицу коэффициентов парной корреляции r_{ij}, $i > j$, характеризующих тесноту линейной связи переменных i и j, перечень которых включает отклик Y (Bmass) и все остальные изучаемые факторы. Ниже главной диагонали симметричной корреляционной матрицы поместим оценки p статистической значимости, которые соответствуют вероятностям ошибочного отклонения нулевой гипотезы H_{ij0}: $|r_{ij}| = 0$. Нетрудно заметить, что данные об ионном составе почвы, исключая содержание бикарбонатов (Carbon), представляют собой однородный комплекс сильно коррелированных переменных, статистически значимо связанных с биомассой травостоя.

	Bmass	Carb	Sulfat	Cl	Ca	Mg	Na	H
Bmass	1	-0.095	-0.330	-0.205	-0.258	-0.310	-0.297	-0.541
Carb	0.2351	1	-0.012	-0.003	-0.072	-0.016	0.059	0.438
Sulfat	< 0.001	0.88	1	0.491	0.820	0.846	0.822	0.346
Cl	0.0096	0.97	< 0.001	1	0.526	0.788	0.741	0.105
Ca	0.001	0.369	< 0.001	< 0.001	1	0.711	0.541	0.270
Mg	< 0.001	0.840	< 0.001	< 0.001	< 0.001	1	0.812	0.242
Na	< 0.001	0.460	< 0.001	< 0.001	< 0.001	< 0.001	1	0.281
H	< 0.001	< 0.001	< 0.001	0.187	< 0.001	0.002	< 0.001	1

В нашем случае мультиколлинеарность приводит к тому, что оценки большинства коэффициентов полной линейной модели регрессии 1-го порядка оказываются статистически незначимыми по t-критерию, хотя F-критерий свидетельствует о значимости всего уравнения в целом:

Предикторные переменные	Показатель VIF	Коэффициенты b_j	Стандартная ошибка S_{bj}	t-критерий	p-значение Pr(>t)
Св. член		195.05	20.3	9.6	**< 0.00001**
H	1.48	-72.39	10.01	-7.22	**< 0.00001**
Carb	1.31	121.9	57.29	2.13	**0.035**
Mg	5.24	-3.01	2.45	-1.232	0.22
Ca	2.15	1.025	1.73	0.594	0.553
Cl	2.99	-0.286	1.54	-0.186	0.853
Na	3.35	-0.122	1.544	-0.08	0.937
Sulfat	Расчет коэффициента не проведен по причине ошибки сингулярности				

Более того, при построении модели с использованием функций lm() или glm() среды R появляется сообщение "матрица сингулярна". Это означает, что численные методы, которые используются при нахождении корней системы нормальных уравнений, не могут найти решение с заданной точностью (ошибка вычислений от итерации к итерации не уменьшается, а возрастает). Статистические характеристики этой и всех остальных моделей приведены в сводной таблице 4.5 в конце следующего раздела.

Обычно рекомендуется исключить из регрессионной модели незначимые коэффициенты, или, выражаясь точнее, выполнить отбор информативного комплекса из q переменных ($q < m$). Модель с настроенными параметрами, доставляющая минимум заданному функционалу качества $L(\boldsymbol{\beta}, \boldsymbol{x})$, называется *моделью оптимальной структуры*. Поскольку глобального оптимума при больших m достичь невозможно, представляется разумным получить в результате расчета для последующего содержательного анализа некоторый набор субоптимальных моделей-претендентов.

Принято считать, что частными критериями качества являются стандартное отклонение для остатков S_e и число отобранных параметров q. Коэффициент детерминации R^2, F-критерий и некоторые другие статистики являются мерами, коэквивалентными S_e, поскольку не меняют предупорядоченность моделей при их

селекции. Использовать для поиска наилучшей модели любой из двух частных критериев S_e или q в отдельности некорректно: минимизация ошибок регрессии S_e является, как правило, "жадным" процессом, стремящимся увеличить параметричность модели $q \to m$, а самая экономная модель, наоборот, имеет естественный минимум при $q = 0$.

Определенный компромисс для поиска компактной модели регрессии, адекватно описывающей результаты наблюдений, доставляют комплексные критерии $L(\beta, \boldsymbol{x}, S_e, q)$, учитывающие обе конкурирующие тенденции. Таковым можно считать приведенный коэффициент детерминации $\overline{R}^2$, однако более взвешенный подход к поиску оптимальной модели реализуется с использованием критерия Акаике, связь которого с априорными вероятностями байесовского подхода описана в работе (Burnham, Anderson, 2002).

Таким образом, когда говорится об оптимальных моделях регрессии, имеются в виду одна или несколько моделей, доставивших субэкстремальные значения выбранному критерию качества (т.е при минимуме AIC или максимуме $\overline{R}^2$). Существует целый арсенал методов, реализующих процедуры селекции признаков оптимальной структуры:

- последовательные или шаговые методы;
- полный перебор всех возможных претендентов;
- использование специальных методов оценивания коэффициентов со штрафом на увеличение их числа (гребневая регрессия, алгоритм Лассо, метод наименьших углов);
- алгоритмы эволюционного поиска: метод модельной закалки (simulated annealing), генетический алгоритм, случайный поиск с адаптацией и др.

Шаговые алгоритмы состоят, как правило, из трех процедур:

- "последовательное включение" (forward) – на первом шаге из всех m признаков выбирается наилучший по заданному критерию; на втором шаге выбирается признак, который в сочетании с первым дает оптимальное решение, и т.д.; процесс заканчивается, когда критерий достиг экстремума и в модель включены q переменных;
- "последовательное исключение" (backward) – на первом шаге перебираются все комбинации из m переменных и исключается наименее информативный признак с точки зрения заданного критерия; эти шаги повторяются, пока критерий не достигнет экстремума или не будут выполнены какие-то иные условия расчета;
- "включение с исключениями" (both) – комбинированный алгоритм, в котором за несколькими шагами включений следуют несколько шагов исключений и т.д.

Использование процедуры stepAIC(…), выполняющей оценку моделей-кандидатов по информационному AIC-критерию, дало для нашего примера следующие результаты:

Предикторные переменные	Коэффициенты b_i	Стандартная ошибка S_{bj}	t-критерий	p-значение Pr(>t)
Св. член	197.3	19.5	10.1	**< 0.00001**
H	-70.96	9.44	-7.52	**< 0.00001**
Carb	114.8	55.7	2.06	**0.0412**
Mg	-2.8	1.105	-2.53	**0.0123**

В отличие от функции полной линейной регрессии, модель на основе информативного комплекса переменных уже является ограниченно мультиколлинеарной, что делает возможным ее интерпретацию с целью предметного "объяснения" причинно-следственных отношений в изучаемом объекте. В частности, можно предположить, что биомасса травянистых растений увеличивается пропорционально снижению высоты и содержания ионов магния и увеличению содержания бикарбонатов. Можно также отметить, что статистические характеристики модели на основе "информативного" набора переменных также существенно улучшились – см. таб. 4.5.

Трудность проблемы формирования оптимального подмножества признаков шаговыми методами обусловлена тем, что после отбрасывания одного предиктора модели соотношение критериев значимостей остальных анализируемых переменных в общем случае изменяется. Поэтому, если в исходной таблице между признаками имеются

сложные корреляционные взаимоотношения, то последовательные процедуры не всегда приводят к результату, достаточно близкому к оптимальному.

Метод регрессии Лассо (LASSO - Least Absolute Shrinkage and Selection Operator) заключается во введении дополнительного регуляризующего слагаемого в минимизируемый функционал, что часто позволяет получать более устойчивое решение. При этом оценки параметров модели $\hat{\beta}$ находятся из условия минимизации

$$\hat{\beta} = \arg\min[\sum_{i=1}^{n}(y_i - \sum_{j=1}^{m}\beta_i x_{ij})^2 + \lambda|\boldsymbol{\beta}|]$$

т.е. достигается некоторый компромисс между ошибкой регрессии и размерностью используемого признакового пространства, выраженного суммой абсолютных значений коэффициентов $|\boldsymbol{\beta}|$. В ходе минимизации некоторые коэффициенты становятся равными нулю, что, собственно, и определяет отбор информативных признаков.

При значении параметра регуляризации $\lambda = 0$, регрессия Лассо сводится к обычному методу наименьших квадратов, а при его увеличении формируемая модель становится все более лаконичной. Оптимальная величина λ находится с использованием кросс-проверки, т.е. ей соответствует минимальная ошибка прогноза $\hat{y}_i$ на экзаменуемых примерах, не участвовавших в построении самой модели.

Для рассматриваемого примера было найдено $\lambda_{опт} = 0.636$ и получена модель регрессии Лассо, по сути идентичная полученной ранее линейной модели lm(), но с чуть большей стандартной ошибкой ($S_e = 61.4$): Y = 74.13 + 36.16Carb - 2.21Mg - 58.7H.

Прямой путь решения задачи оптимально состава признаков заключается в полном переборе всех возможных C_m^k сочетаний переменных ($k = 0, 1, \ldots, m$). Воспользуемся для этого функцией glmulti() из одноименного пакета среды R, где для отбора кандидатов будем использовать тот же AIC-критерий. В классе моделей 1-го порядка при 7 независимых переменных число претендентов составляет $(2^m - 1) = 127$ вариантов, что вполне реализуемо полным перебором. Три лучшие модели со всеми статистически значимыми коэффициентами имеют вид:

Y = 197.3 + 114.8Carb - 2.797Mg - 70.9H (AIC = 1763.6, $s_{CV} = 62.3$);
Y = 195.1 + 120.8Carb + 1.05 Ca - 3.44Mg - 72.2H (AIC = 1765.2, $s_{CV} = 62.5$);
Y = 194.9 + 128.0Carb - 1.92 Cl - 75.2H (AIC = 1765.4, $s_{CV} = 63.0$),

т.е. в нашем примере полный перебор соответствовал результатам шаговой процедуры.

Принципиально иной подход к нахождению оптимальной регрессии основан на внешних критериях, учитывающих ошибку прогноза на экзаменуемых примерах, не участвовавших в построении самой модели. Такими критериями могут быть средние характеристики качества моделей, подвергаемых процедуре кросс-проверки:

- среднеквадратичная ошибка $s_{CV} = \left[\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\right]^{0.5}$ или среднее абсолютное отклонение $d_{CV} = \sum_{i=1}^{n}|y_i - \hat{y}_i|/n$, где $\hat{y}_i$ - прогноз при скользящем контроле;
- коэффициент детерминации $R^2_{CV} = 1 - s^2_{CV}/s^2_y$, где s^2_y – оценка дисперсии зависимой переменной.

В некоторых случаях внутренние и внешние критерии приводят к различным результатам, что вызывает неопределенность статистических выводов. Например, модель

Y = 220.2 - 3.1Mg - 62.1H (AIC = 1765.9, $s_{CV} = 62.1$),

является лучшей при кросс-проверке, но не является оптимальной по AIC-критерию.

В общем случае оценка доверительных интервалов параметров множественной регрессии параметрическими методами представляет собой нетривиальную проблему и основывается на сложных аналитических формулах, учитывающих ковариацию между переменными модели. Эта задача упрощается с использованием процедур бутстрепа и рандомизации, техника реализации которых, описанная нами в случае парной регрессии в разделах 3.3-3.4, в равной степени может применяться и для многомерного случая.

Например, помощью функции boot(…) для наилучшей регрессионной модели на основе трех факторов нами были рассчитаны стандартные ошибки и интервальные оценки различных критериев качества. В частности, для коэффициента детерминации $R^2 = 0.335$ получены: бутстреп-смещение $\Delta_{R2} = 0.02$, стандартная ошибка $se_{R2} = 0.0418$ и 95% доверительные интервалы методом BCa $CI_R^2 = \{0.259 \div 0.408\}$.

Аналогично были построены статистические распределения коэффициентов модели b_j, найдены их доверительные интервалы (BCa), уточнены бутстреп-смещения и стандартные ошибки:

Предикторные переменные	Коэффициенты b_i	Смещение $b_i - \bar{b}_j^*$	Бутстреп-ошибка se_{boot}	95% доверительные интервалы
Св. член	197.3	0.399	20.33	
H	-70.96	-0.153	13.26	-93.81 ÷ -40.92
Carb	114.8	-1.766	89.12	-88.6 ÷ 256.9
Mg	-2.8	0.037	0.878	-4.404 ÷ -0.982

Поскольку доверительные интервалы для параметра при переменной Carb включают значение 0, нет оснований считать этот коэффициент статистически значимым, что косвенно подтверждается результатами кросс-проверки. Для множественной регрессии дополнительно может быть изучен характер взаимного влияния и совместного распределения коэффициентов моделей для произвольных пар индивиуальных факторов. Например, на рис. 4.4 показано поле корреляции бутстреп-оценок коэффициентов для двух переменных Carb и высоты H, а также проведены доверительные эллипсы интервалов с заданными вероятностями 0.9, 0.95 и 0.99, которые основаны на вычислении робастных оценок ковариационной матрицы коэффициентов (Fox, 2002).

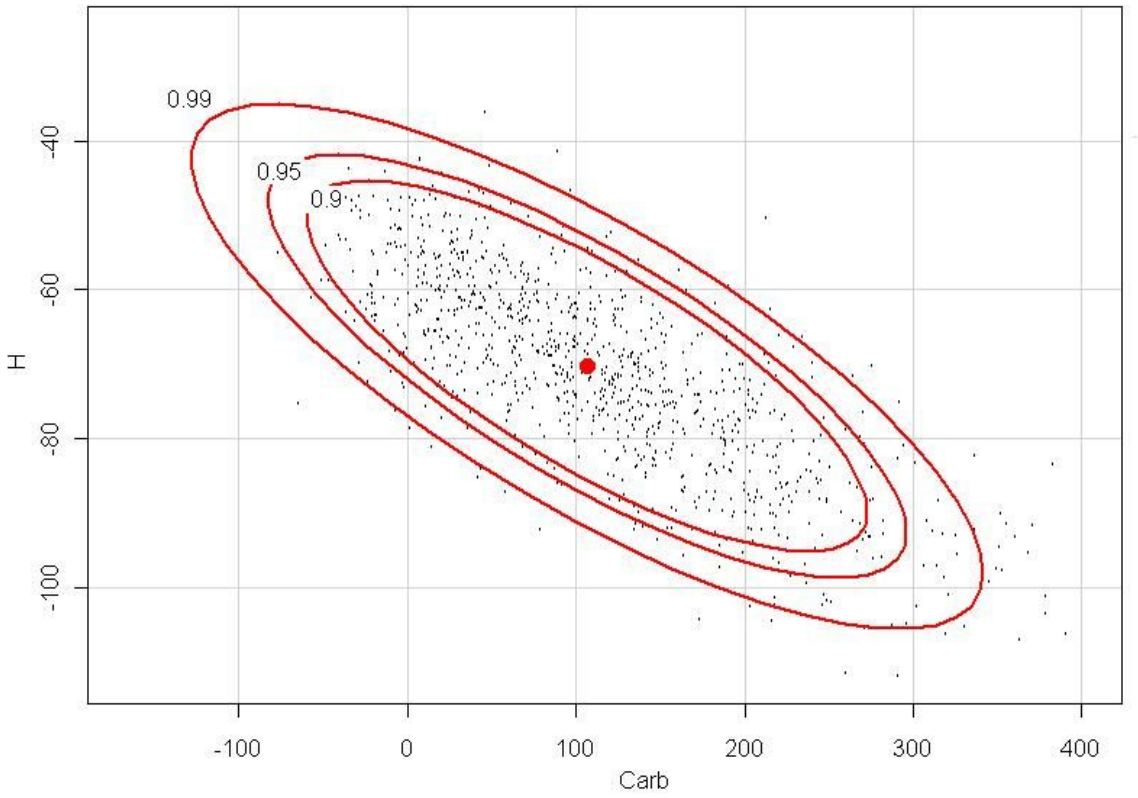


Рис. 4.4. Диаграмма рассеяния бутстреп-оценок коэффициентов регрессии для переменных, соответствующих содержанию бикарбонат-иона Carb и высоты пробной площадки H

К разделу 4.5:

```
# Загрузка данных из предварительно подготовленного двоичного файла
load("Fito.RData")  ;  library(glmulti) ; library(boot) ; library(MASS) ; library(car)
```

```
corsig <- function(x){  # Функция расчета корреляционной матрицы с p-значениями
x <- as.matrix(x) ; R <- rcorr(x)$r ; p <- rcorr(x)$P
ipa <- lower.tri(p, diag = FALSE) ; R[ipa] <- p[ipa] ; return (R) }
write.table(round(corsig(Fito),4),"clipboard",sep="\t")  # Вывод в буфер обмена
#  Получение полной линейной модели 1-й степени
RegModel.2 <- lm(Bmass~Ca+Carb+Cl+H+Mg+Na+Sulfat, data=Fito) ; summary(RegModel.2)
vif(lm(Bmass~Ca+Carb+Cl+H+Mg+Na, data=Fito))  #   Расчет VIF-статистики
stepAIC(RegModel.2,data=Fito)  # Оптимизация модели с применением шаговой процедуры
library(lars)  #  Получение регрессионной модели Лассо
m.lasso <- lars(as.matrix(Fito[,2:8]),Fito[,1]) ; plot(m.lasso)
# Кросс-проверка
r <- cv.lars(as.matrix(Fito[,2:8]),Fito[,1]) ; bestfrac <- r$index[which.min(r$cv)]
coef.lasso <-
      predict(m.lasso,as.matrix(Fito[,2:8]),s=bestfrac,type="coefficient",mode="fraction")
coef.lasso  # Получение коэффициентов модели
y.pred.lasso <-
predict(m.lasso,as.matrix(Fito[,2:8]),s=bestfrac,type="fit",mode="fraction")$fit
summary(sqrt((y.pred.lasso - Fito[,1])^2))  #   Сумма квадратов отклонений модели
# Оптимизация модели 1-й степени путем полного перебора
prednames <- names(Fito)[2:8] ; glmulti("Bmass",xr=prednames,data=Fito,family =
                                        gaussian,level=1)
sqrt(cv.glm(Fito,  glm(Bmass~Carb+H+Mg, data=Fito))$delta) # Получение ошибки кросс-проверки
summary(m2 <- glm(Bmass~H+Mg, data=Fito)) ; sqrt(cv.glm(Fito, m2)$delta)
# Бутстреп-анализ коэффициентов уравнения регрессии и других статистик модели
# функция, возвращающая вектор коэффициентов
bootF <- function(data, indices){data <- data[indices,]
      mod <- lm(Bmass ~ Bmass ~ 1 + Carb +H + Mg, data=data) coefficients(mod) }
Fito.boot <- boot(Fito, bootF, 1000) ; Fito.boot
boot.ci(Fito.boot,  index = 2 , type= type = c("norm", "basic", "perc", "bca")) # Carb
data.ellipse(Fito.boot$t[,2], Fito.boot$t[,3], xlab="Na", ylab="H",
                                  cex=.3, levels=c(.9, .95, .99), robust=T)
#  Бутстреп-оценивание коэффициента детерминации
rsq <- function(data, indices){data <- data[indices,]
      mod <- lm(Bmass ~ 1 + Carb +H + Mg, data=data) ; return(summary(mod)$r.square)}
Rsq.boot <- boot(Fito, rsq, 1000)  ; boot.ci(Rsq.boot,   type="bca") # R2
```

4.6. Селекция моделей: генетический алгоритм и случайный поиск с адаптацией

Выбор исходного пространства переменных, критериев качества конструируемых статистических моделей и выражений для регрессионных функций во многом зависят от характера прикладных задач. Многомерный регрессионный анализ в общем случае может служить различным целям исследований:

1. *Верификация* теоретических моделей. При этом предметная модель жестко обуславливает модель статистическую, предписывая ей определенные спецификации, включающие в себя функциональную структуру и требуемые переменные.

2. Обоснование предметных гипотез и "*объяснение* механизмов". Отталкиваясь от данных мониторинга и пользуясь аппаратом проверки статистических гипотез, делается вывод о наличии в данных элементов внутренней детерминированной структуры (степени взаимного влияния между подмножествами переменных, группировки объектов в кластеры и т.д.). Модельные изыски при этом обычно минимальны – модели достаточно лаконичны и служат для подтверждения тех или иных теоретических предположений.

3. Поиск и выделение в данных нетривиальных конструкций или характерных шаблонов (pattern), что обычно объединяется под названием *data mining*. Поскольку речь идет о конструировании новых знаний, задача оценки статистической достоверности ставится достаточно строго, но ограничения на спецификацию моделей минимальны.

4. Построение *прогнозов*. Для выбора прогнозирующих моделей используются алгоритмы и внешние критерии, связанные только с качеством подгонки под данные (goodness of fit). Сами модели бывают очень сложны или вообще не иметь явной

параметрической формы (например, нейронные сети, ядерные функции, модели МГУА или random forrest), поэтому классический статистический анализ значимости факторов и оценку взаимосвязей между ними в полной мере часто провести невозможно.

Одной из задач data mining является поиск статистически значимых множественных взаимодействий между факторами при их совместном влиянии на отклик *y*. В общем случае аддитивные модели множественной регрессии для *m* независимых переменных могут иметь различный состав компонент, простирающийся от линейного уравнения 1-го порядка (4.2) до обобщенного полинома Колмогорова-Габора:

$$y = b_0 + \sum_{i=1}^{m} b_i x_i + \sum_{i=1}^{m}\sum_{j=1}^{m} b_{ij} x_i x_{ij} + \sum_{i=1}^{m}\sum_{j=1}^{m}\sum_{k=1}^{m} b_{ijk} x_i x_j x_k + \ldots .$$

Если при $m = 7$ ограничиться только парными произведениями переменных, то необходимо дополнительно рассчитать $m(m\text{-}1)/2 + m = 28$ коэффициентов b_{ij}. Однако число возможных комбинаций моделей, которые можно составить из этих термов, будет равно $2^{\text{choose}(7,2)+7} = 268.4$ миллионов вариантов, что потребует слишком большого объема вычислений (здесь choose(*m*, 2) – функция R, возвращающая биномиальный коэффициент). Поэтому полный перебор моделей с парными взаимодействиями практически возможен лишь при *m* не больше 5. В иных случаях для поиска вариантов моделей, доставляющих оптимум заданному критерию, можно реализовать элегантный автоматизированный подход с использованием *генетического алгоритма*, который можно считать "интеллектуальной" формой проб и ошибок в семействе методов Монте-Карло.

Генетический алгоритм, позаимствованный у природных аналогов, базируется на двух основных следствиях эволюционной теории, сформулированных Чарльзом Дарвиным в работе «Происхождение видов»:

- естественный отбор является движущей и направляющей силой эволюции, что предполагает некоторый механизм выделения самых сильных и полезных экземпляров (решений, структур, особей, алгоритмов);
- необходимо наличие некоторых степеней свободы эволюционного процесса в виде изменчивости объектов, т.е. возможности генерации принципиально новых структур искомых объектов (перечисление то же) в виде непрекращающейся последовательности "проб и ошибок".

Именно эти принципы отбора наилучших объектов являются ключевой эвристикой всех эволюционных математических методов, позволяющих часто уменьшить время поиска решения на несколько порядков по сравнению со случайным поиском. Механизм естественного отбора связывается здесь с принятым критерием оптимальности $IC(\boldsymbol{\beta}, \boldsymbol{x})$, определяющим сравнительную ценность произвольного варианта эволюции, а изменчивость вносится путем специальных модификаций фрагментов бинарного кода.

Генетический алгоритм был разработан Дж. Холландом (Holland) в 1975 году в Мичиганском университете и отличается от различных эвристических процедур и обычных случайных методов Монте-Карло тем, что поиск оптимального решения развивается не сам по себе, а с учетом предыдущего опыта. В каноническом виде алгоритм описывается следующей совокупностью шагов (Goldberg, 1989):

1) Задается критерий оптимальности $IC(\boldsymbol{\beta}, \boldsymbol{x})$, определяющий эффективность каждой произвольной комбинации признаков. В нашем случае это могут быть различные формы информационных критериев Акаике AIC или Байеса BIC, оценивающих качество регрессионной модели. Формируемое решение кодируется как вектор $\boldsymbol{x}^0$, который называется хромосомой и соответствует битовой маске, т.е. двоичному представлению набора исходных переменных. Элементами хромосом являются части вектора или "гены", изменяющие свои значения в определенных позициях – "аллелях" или "локусах".

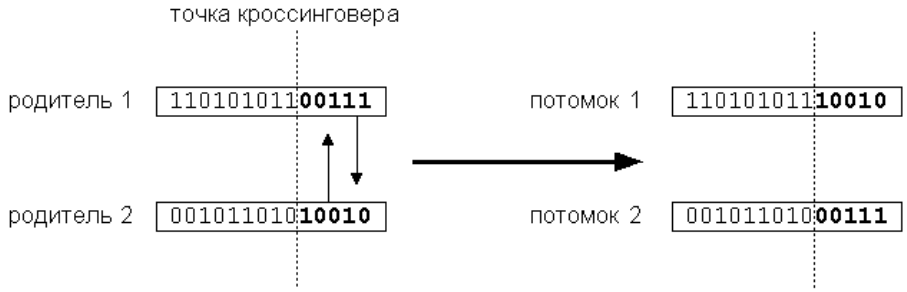
2) Случайно или в соответствии с определенными ограничениями инициализируется исходная "популяция" $\Pi^0(x_1^0..x_\lambda^0)$ потенциальных решений, состоящая из некоторого количества хромосом λ, которое является параметром алгоритма.

3) Каждой хромосоме x_i^0, $i = 1,...,\lambda$ в популяции присваиваются две оценки: значение эффективности $\mu(x_i^0)$ в соответствии с вычисленной функцией оптимальности и вероятность воспроизведения $P(x_i^0)$, которая зависит от перспективности этой хромосомы. Например, эта вероятность будет больше, чем выше $w_i = e^{-(IC_i - IC_{best})}$, где IC_{best} – лучший информационный критерий в текущей совокупности моделей.

4) В соответствии с вероятностями воспроизведения $P(x_i^0)$ создается новая популяция хромосом, причем с большей вероятностью воспроизводятся наиболее перспективные элементы. Хромосомы производят потомков, используя три операции: кроссинговера (или рекомбинации), простой мутации и иммиграции.

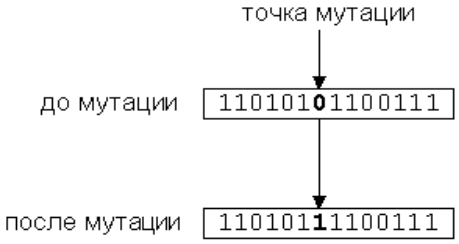
5) Процесс останавливается, если получено удовлетворительное решение, либо если исчерпано все отведенное на эволюцию время. Если процесс не окончен, то вновь повторяется воспроизведение новой популяции и поиск субоптимальных моделей.

Оператор *кроссинговера* на шаге 4 производит скрещивание хромосом и обмен генетическим материалом между родителями для получения потомков. Этот оператор служит для улучшения существующих областей пространства (*эволюционное приспособление*). Простейший одноточечный кроссинговер производит обмен частями, на которые разбивает хромосому точка кроссинговера, выбираемая случайно:



Двухточечный кроссинговер обменивает кусок строки, попавшей между двумя точками. Предельным случаем является равномерный кроссинговер, в результате которого все биты хромосом обмениваются с некоторой вероятностью.

Оператор *мутации* нужен для расширения пространства поиска ("*эволюционное исследование*") и предотвращения невосстановимой потери бит в аллелях. Он применяется к каждому биту хромосомы с небольшой вероятностью ($p \approx 0.001$), в результате чего бит (аллель) изменяет значение на противоположный.



Оператор иммиграции заключается в принудительном назначении случайному аллелю значения 0 или 1 в равной вероятности. Он обеспечивает наибольшие возмущения в структуре популяции, что позволяет избежать "рыскания" вблизи локального оптимума.

Для управления процессом селекции моделей 2-го порядка в примере, представленном в разделе 4.6, мы задавали следующие значения входных параметров функции glmulti(…), принимаемые, в основном, по умолчанию:

- размер популяции λ определялся значением popsize = 100;
- правила остановки устанавливали параметры deltaM = 0.05, deltaB = 0.05 и conseq = 5;

° интенсивностью кроссинговера управлял параметр sexrate = 0.1, иммиграции – imm = 0.3; вероятность мутаций была задана mutrate = 0.01.

После смены 420 поколений популяций, каждая из которых насчитывала 100 претендентов, были получены модели с парными произведениями, существенно лучшие по AIC-критерию, чем модели 1-го порядка (см. профиль изменения AIC на рис. 4.5).

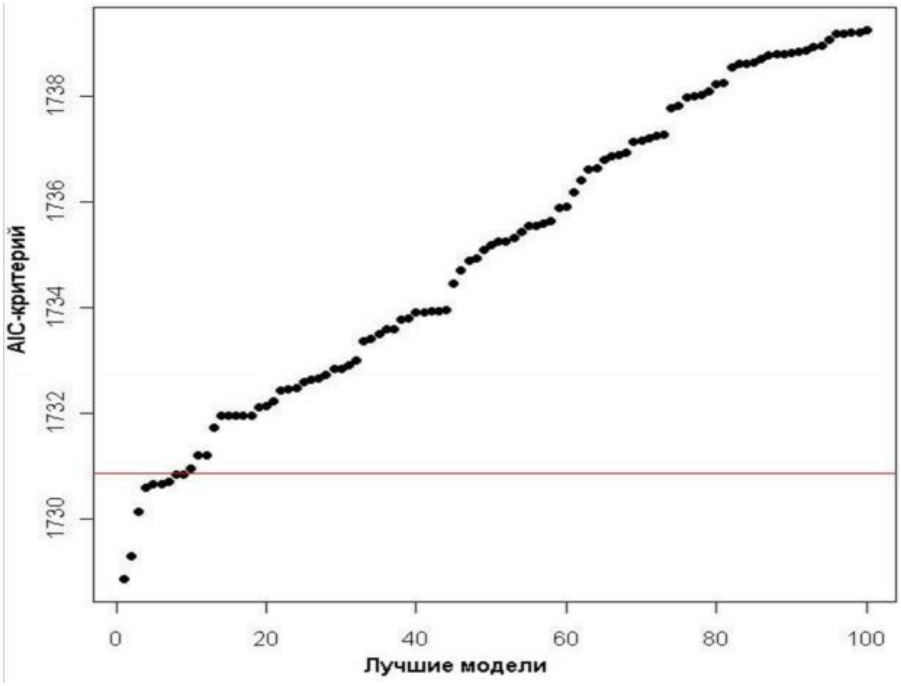


Рис. 4.5. Профиль информационного критерия Акаике для 100 лучших моделей 2-го порядка, учитывающих парные взаимодействия факторов

Для содержательной интерпретации результатов регрессионного анализа интересно рассмотреть частоты вхождения отдельных переменных в десятку субоптимальных моделей:

	Na	Carb	H	Cl	Ca	Mg	Sulfat
Как главный фактор	10	10	10	10	2	1	1
В парных взаимодействиях	22	18	18	3			

Этот показатель является дополнительной оценкой степени влияния факторов.

Регрессионная модель, учитывающая парные произведения предикторов и доставляющая минимум AIC-критерию представлена в табл. 4.4 (пары предикторов показаны с использованием символа ‘:’, как это принято в нотации моделей R). Представленная модель существенно превышает линейные модели 1-го порядка по всему комплексу основных характеристик (см. табл. 4.5). Интересно, что влияние содержания иона натрия, ранее единодушно признаваемое незначимым, в моделях с взаимодействиями заняло место, достойное самого пристального внимания.

Таблица 4.4. Статистический анализ коэффициентов модели, учитывающей парные произведения; символ ':' связывает факторы, для которых рассчитывается эффект взаимодействия

Предиктор-ные переменные	Параметрические оценки				95% доверительные интервалы бутстрепа (BCa)
	Коэффи-циенты a_i	Стандартная ошибка S_{bj}	*t*-критерий	*p*-значение Pr(>*t*)	
Св. член	198.1	45.8	4.3	**< 0.00001**	
Na	-16.79	3.7	-4.5	**< 0.00001**	-23.46 ÷ -9.63
H	-106.7	24.5	-4.35	**< 0.00001**	-166.7 ÷ -50.7
Na:H	11.3	1.69	6.68	**< 0.00001**	6.12 ÷ 16.11
Carb	502.7	147.5	3.4	**0.00084**	-188.1 ÷ 1098.5
Carb:Na	-34.28	10.9	-3.11	**0.0021**	-59.19 ÷ -6.22
Carb:H	-97.8	64.12	-1.526	0.129	-304.78 ÷ 147.51
Cl	2.85	1.31	2.16	**0.031**	0.913 ÷ 4.858

Другой проблемой идентификации адекватных регрессионных моделей является проверка справедливости линейной спецификации модели (4.2). Существующие в природе количественные соотношения между экологическими процессами в принципе являются нелинейными, а приведение их к линейной форме – лишь одна из возможностей удобной аппроксимации, которая может оказаться весьма грубой. Для оценки степени нелинейности можно, например, построить обобщенную аддитивную модель GAM (подробности см. далее в разделе 4.7) и рассчитать значения *эффективных степеней свободы* (EDF – effective degrees of freedom) нелинейных сглаживающих функций для каждой из независимых переменных. Значения EDF, полученные с использованием функции gam() пакета mgcv (Wood, 2006), для рассматриваемого примера равны:

H	Na	Ca	Carb	Mg	Cl	Sulfat
5.96	2.25	1.38	1.13	0.88	0.84	0.53

Значения оценок степеней свободы, превышающие 1.5, свидетельствуют о существенной нелинейности связи этих переменных с откликом – см. рис. 4.6.

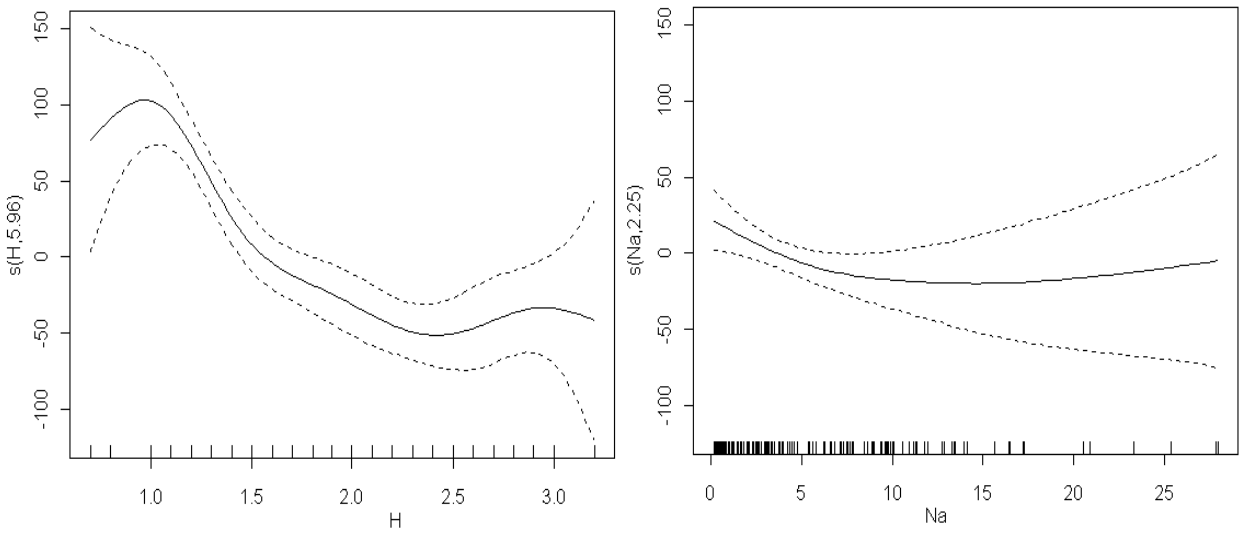


Рис. 4.6. Вид нелинейных сглаживающих функций s(…) зависимости биомассы растительности от высоты пробной площадки H и содержания ионов Na в почве

Другой простой способ провести тест на функциональную форму – это добавить в правую часть уравнения (4.2) различные нелинейные члены и проверить, как изменится при этом принятый критерий качества. Очень часто бывает, что недоопределенность линейной модели при этом статистически значимо уменьшается. Для расширения признакового пространства за счет нелинейных преобразований исходных переменных необходимо определить конечное множество порождающих функций (Стрижов, Крымова, 2010).

Для каждой из 7 исходных переменных, определяющих ионный состав почвы в примере [П3], выполним формирование новых векторов с использованием 6 порождающих функций: $\ln X$, $\sqrt{\boldsymbol{x}}$, $1/X$, X^2, e^X, и $e^{1/X}$. Включив векторы с трансформированными данными в исходную таблицу, получим матрицу 50×159, один столбец которой содержит зависимую переменную – биомассу травянистого покрова. Требуется найти модель оптимальной структуры, которая доставляет минимальное значение функционалу качества $L(\boldsymbol{\beta}, \boldsymbol{x}, S_e, q)$ и состоит из небольшого подмножества наиболее информативных членов. Не лишним будет отметить, что общее число возможных вариантов моделей равно $(2^{49} - 1) = 5.63 \cdot 10^{14}$ и это дало повод отказаться от расчета функциям среды R, реализующим генетический алгоритм и регрессию Лассо.

Алгоритм СПА (Лбов, 1981; Загоруйко, 1999) выполняет адаптивный случайный поиск наиболее информативного подмножества переменных. При этом производится "поощрение" и "наказание" отдельных предикторов в зависимости от результатов построения набора моделей на различных опорных подмножествах признаков. Процесс стартует с присвоения элементам вектора **W** весов $\{w_1, w_2, \ldots, w_m\}$ равных значений $w_i = 1/m$, т.е. все признаки с равной вероятностью могут претендовать для отбора в модель. Далее выполняется следующая последовательность действий:

- проводится T ($T \approx 100$) серий испытаний с построением моделей различной сложности $\{1, 2, \ldots, q\}$, $m \geq q$;
- в каждой серии формируется по N ($N \approx 100$) наборов переменных, причем отбор признаков осуществляется случайно, но с учетом весов вероятностей w_i ; по этим наборам осуществляется построение регрессионных моделей, после чего выделяется наилучшая и наихудшая модели из N в соответствии с заданным критерием качества $L(\boldsymbol{\beta}, \boldsymbol{x}, S_e, q)$;
- после завершения серии веса w_i пересчитываются: для всех признаков, попавших в "хорошую" модель выполняется "поощрение" $w_i = w_i + \Delta w_p$, а для признаков "плохой" модели – "наказание" $w_i = w_i - \Delta w_h$; при этом соблюдается условие нормировки $\sum w_i = 1$;
- предъявляется модель, показавшая лучшие результаты во всех сериях испытаний.

На данных рассматриваемого примера алгоритм СПА показал хорошую скорость сходимости, однако при использовании информационного критерия Акаике он так и не смог преодолеть обширного локального минимума функционала в районе AIC = 1733. Тем не менее, был сформирован набор вполне интерпретируемых и лаконичных моделей, имеющих лучшие характеристики, чем любая линейная модель 1-го порядка, например:

$$Y = 1527 + \{183H^2 - 1848 \ln(H) - 30.2e^H - 1463/H\} + 142(Carb)^{0.5} - 23.8(Na)^{0.5} - 30.4/Sulfat.$$

Статистические характеристики модели приведены в табл. 4.5.

На этом фоне стандартный алгоритм "включений с исключениями" дал несколько обескураживающие результаты: максимум приведенного коэффициента детерминации R^2 (см. табл. 4.5) был получен для модели, включающей 31 переменную(!) из 49, что делает ее мало пригодной для какой-либо содержательной интерпретации. Эта неудача неизбежно приводит к мысли, что R^2, AIC-критерий или любая другая внутренняя статистика, основанная на стандартных отклонениях для остатков, не всегда являются эффективным инструментом поиска субоптимальных моделей в сложных случаях.

К значительно более позитивным результатам привело использование метода “Forward selection” (Miller, 1990), основанного на внешних критериях, получаемых процедурой кросс-проверки. Алгоритм селекции моделей осуществляет перебор всех уровней сложности моделей $\{1, 2, \ldots, q\}$ и для каждого уровня ищется модель, оптимальная с точки зрения критериев s_{CV}, d_{CV} или R^2_{CV}. Лучшей принимается модель при той размерности переменных, которая доставляет максимум принятому внешнему критерию – см. рис. 4.7.

q	R^2_{CV}	Вид модели
1	0.26	Y = -38.59 + 210/H
2	0.31	Y = -17 + 198/H - 12.7ln(Na)
3	0.32	Y = -94.82 + 218/H - 11.4 ln(Na) + 114(Carb)$^{0.5}$
…		…
7	0.42	Y = 430 + {571H - 1503ln(H) - 309.8e^{H} } + 145.5(Carb)$^{0.5}$ - 19.5ln(Na) - - 21.1/Cl - 15ln(Cl)

Рис. 4.7. Изменение коэффициента детерминации при кросс-проверке R^2_{CV} в зависимости от размерности подпространства переменных, используемых при построении моделей; приведены модели, признанные лучшими на некоторых уровнях

Модель оптимальной сложности при q = 7 включает все четыре фактора, что и лучшая модель с парными взаимодействиями (табл. 4.4), однако нелинейный характер влияния факторов учитывается здесь за счет функциональных преобразований исходных переменных. Коэффициенты модели оказались статистически значимыми кроме терма 15ln(Cl) с p = 0.085. Однако исключать этот член нет объективных оснований, поскольку он вносит свой определенный вклад в поддержку качества предсказания. Представляется разумным сформулировать следующее общее правило: "допустимо включение в модель термов со статистически незначимыми коэффициентами, если при этом улучшаются статистики кросс-проверки". Можно также отметить, что модель с q = 7 оказалась также наилучшей и внутреннему информационному критерию качества AIC – см. табл. 4.5.

Таблица 4.5. Сводная таблица основных статистических характеристик моделей регрессии

Модель	Использован-ный алгоритм	Число пре-дикторов	S_e ошибка регрессии	R^2 детерми-нации прив.	F-статистика	AIC-критерий
Линейная	Полная	7	61.36	0.3209	13.4	1769.2
Линейная	Шаговый	3	60.8	0.332	**27.19**	1763.6
Полином	Генетический	7	53.89	0.4761	21.51	1728.9
Нелинейная	Шаговый	31	**50.4**	**0.541**	7.0	1728.3
Нелинейная	СПА	7	54.6	0.462	20.38	1733.1
Нелинейная	Кросс-проверка	7	53.2	0.488	22.57	**1725**

Мы далеки от надежды, что представленные модели будут с энтузиазмом восприняты специалистами по экологии растительных сообществ. Сформировать коллекцию моделей – это далеко не конец обработки данных, а только самое ее начало. Чтобы получить содержательно интерпретируемую модель, необходимо выполнить ряд важных операций (Цейтлин, 2007):

° выделить область определения анализируемой модели путем окаймления поверхностями, зависящими от характера распределения факторов,

- сравнить "силу" влияния на отклик отдельных групп базисных функций для различных интервалах их изменчивости,
- исследовать степень коллинеарности переменных и характер распределения остатков модели и др.

Но это – тема особого разговора, а мы стремились на этом примере только показать сложность и неоднозначность решения проблемы отбора значимых параметров, которая существует в различных областях знаний под названием Dimensionality Reduction (уменьшение размерности).

К разделу 4.6:

```
# Загрузка данных из предварительно подготовленного двоичного файла
load("Fito.RData") ; prednames <- names(Fito)[2:8] ; library(glmulti) ; library(MASS) ;
library(mgcv)
# Получение субоптимальных моделей с взаимодействиями на основе генетического алгоритма
obga <- glmulti("Bmass",xr=prednames,data=Fito,family = gaussian, level=2, method="g",
marginality=TRUE, conseq=5,popsize=100,mutrate=1e-2)
BestFitoModel <- lm(formula = Bmass ~ 1 + Carb + Cl + Na + H + Na:Carb + H:Carb + H:Na,
             data = Fito) ; summary(BestFitoModel)
#  Оценка эффективных степеней свободы сглаживающих функций модели GAM
modGAM = gam(Bmass~s(Ca)+s(Carb)+s(Cl)+s(H)+s(Mg)+s(Na)+s(Sulfat), data=Fito)
summary(modGAM) ; anova(modGAM) ; plot(modGAM)
# Выполение нелинейных преобразований и расширение исходной таблицы
DF1 <- apply( Fito[,2:8], 2,  log); colnames(DF1) <- paste("ln_",colnames(DF1), sep="")
DF2 <- apply( Fito[,2:8], 2,  sqrt); colnames(DF2) <- paste("sqr_",colnames(DF2), sep="")
DF3 <- apply( Fito[,2:8], 2,  exp); colnames(DF3) <- paste("exp_",colnames(DF3), sep="")
DF4 <- apply( Fito[,2:8],2,function(x) x*x)
colnames(DF4) <- paste(colnames(DF4),"_2", sep="")
DF5 <- apply( Fito[,2:8],2,function(x) 1/x)
colnames(DF5) <- paste("g1_",colnames(DF5), sep="")
DF6 <- apply( Fito[,2:8],2,function(x) exp(1/x))
colnames(DF6) <- paste("e1_",colnames(DF6), sep="")
Fito_R <- cbind(Fito, DF1, DF2, DF3, DF4, DF5, DF6)
# Поиск оптимальной модели с применением шаговой процедуры
Model.Full <- lm(Bmass~., data=Fito_R) ; summary(Model.Full); AIC(Model.Full)
stepAIC(Model.Full,data=Fito_R,steps = 5000,trace = FALSE)
#   ------------- Алгоритм СПА - Случайный поиск с адаптацией  ------------------
#  Функция:  Построение линейной модели по заданной комбинации признаков
estim.lm <- function (df, mask) {
       fml <- as.formula(paste(names(df)[1],"~.")) ; df_sel <- df[,c(1,mask)]
       return(lm(fml,df_sel)) }
#  Функция:  Адаптация методом "Поощрение-наказание"
adapt.lm <- function (probsVec,h, b.mask, w.mask) {
       subVec = min(h, probsVec[w.mask]) ; probsNew <- probsVec
       probsNew[w.mask] <- probsNew[w.mask] - subVec
       probsNew[b.mask] <- probsNew[b.mask] + mean(subVec); return (probsNew)
}
#  Функция:  Случайный поиск с адаптацией
rsa = function (df, maxNF=min(10, ncol(df)-1), d=5 , Tit = 20, N = 20, h=0.05) {
# Аргументы:  maxNF - максимальное число переменных в модели
# d — количество шагов после нахождения оптимума ; Tit — количество итераций
# N — количество генерируемых наборов на каждой итерации
# h — скорость адаптации (шаг изменения вероятностей)
#  Результат - объект лучшей модели
    n <- ncol(df)-1 ; p  <- as.numeric(rep(1/n, n)) ;  best.nF = 0
    for (nF in sample.int(maxNF,maxNF,T)) {
         for (iIt in 1:Tit) {
            test.mask = sort(unique(sample.int(n, nF, T, p) + 1))
            test.q = AIC(estim.lm(df, test.mask))
            if (best.nF == 0) {
              best.nF <- nF ; best.mask <- test.mask ; best.q <- test.q
              worst.mask <- test.mask ; worst.q <- test.q }
```

```
            else {  if (best.q > test.q) { best.nF <- nF ; best.mask <- test.mask
                                                  best.q <- test.q }
                  else if (test.q > worst.q) { worst.mask <- test.mask ; worst.q <- test.q } }
          }
        if (nF - best.nF > d)  break
        if  (h > 0)  p <- adapt.lm(p, h, best.mask-1, worst.mask-1)
   }
   return (estim.lm(df, best.mask)) }
#  Выполнение расчетов алгоритмом СПА
best.lm <- rsa(Fito_R, Tit = 100, N = 100,h=0.15) ; summary(best.lm) ; AIC(best.lm)
#  Расчеты по алгоритму "Forward selection" с использованием кросспроверки
library(FWDselect)
qob =qselection(x=Fito_R[,2:50],y=Fito_R[,1],qvector=c(1:14),method="lm",criterion="R2")
plot(qob)  #  Вывод графика зависимости внешнего критерия от сложности модели
selection(x=Fito_R[,2:50],y=Fito_R[,1], q=7, criterion = "R2", method = "lm",
                                                    family = "gaussian")
RegModel.7 <- lm(Bmass~ln_H+ln_Na+sqr_Carb+ln_Cl+H+e1_H+g1_Cl, data=Fito_R)
summary(RegModel.7) ; AIC(RegModel.7)
```

4.7. Процедуры сглаживания и обобщенные аддитивные модели

Как обсуждалось в предыдущем разделе, основная цель регрессионного анализа состоит в осуществлении разумной (т.е. адекватной поставленной задаче) аппроксимации математического ожидания отклика $E(Y|x_1,x_2,\ldots,x_m)$ по обучающей выборке с помощью неизвестной функции регрессии $f(\ldots)$. Если в центре внимания находится попытка "объяснить" внутренние механизмы изучаемого явления и оценить при этом сравнительную роль отдельных факторов, то используются параметрические методы регрессионного анализа. Такой подход предполагает, что искомая модель имеет некоторую предписанную ей функциональную форму и полностью описывается конечным набором параметров. Типичным примером параметрической модели является полиномиальная регрессия, параметрами которой являются вектор коэффициентов β при независимой переменной и случайная ошибка модели σ_Y. При этом молчаливо предполагается, что на всем диапазоне изменения $\boldsymbol{x}$ можно найти такое приближение данных единой моделью, ошибка аппроксимации которой будет приемлема в практических целях. Естественно, что неверная спецификация функциональной формы может привести к серьезным, а часто и непредсказуемым искажениям при инференции.

Поскольку каждый предиктор X принимает конечное множество значений, мы всегда проводим *интерполяцию*, т.е. мысленно заполняем промежуточные значения между двумя последовательными выборочными значениями x. Мы также должны принять во внимание, что каждая реализация отклика y_i – выборка из условного распределения $Y|X = x_i$, которая не вполне равна E [Y|X = x_i] и является некоторой *экстраполяцией*. Таким образом, любой способ оценки функции регрессии выполняет интерполяцию, экстраполяцию и аппроксимацию, однако каждый метод вычислений расставляет свои акценты в решении этих задач

Выборочное среднее и простая линейная регрессия – примеры линейных сглаживающих фильтров (smoothers), которые являются оценками функции регрессии для произвольного текущего значения x в следующей форме:

$$\hat{y}(x) = \sum_{i=1}^{n} y_i w(x_i, x), \tag{4.3}$$

где $w(\ldots)$ – некоторая заданная функция расчета весовых коэффициентов, оцениваемых по набору выборочных значений x_i.

Многие классические статистические формулы является частным случаем выражения (4.3). Например, приняв $w(x_i, x) = 1/n$ для x и всех x_i, получим в итоге простое среднее для y. Оценки отклика по методу k-ближайших соседей основаны на

взвешивающей функции $w(x_i, x) = 1/k$, если между x_i и x находится менее k членов, и $w(x_i, x) = 0$, в противном случае. Наконец, простая модель линейной регрессии (без свободного члена) приобретает привычный вид, если положить $w(x_i, x) = (x_i / ns_x^2)\, x$, поскольку $\hat{y}(x) = \hat{b}x = x\sum_{i=1}^{n} x_i y_i / \sum_{i=1}^{n} x_i^2 = \sum_{i=1}^{n} y_i (x_i / ns_x^2)x$.

Если конкретный набор оцениваемых параметров модели значительного самостоятельного интереса не представляет, а важна эффективная интерполяция значений y с минимальной ошибкой, то прекрасной альтернативой классической регрессии является непараметрическое *сглаживание*[7]. При таком методе аппроксимации не производится втискивание данных в "прокрустово ложе" фиксированной параметризации, а сама модель регрессии имеет динамический функциональный характер, подстраиваемый под текущие значения предикторов.

Процедуры сглаживания и другие методы непараметрической аппроксимации предоставляют в настоящее время целый набор гибких и эффективных средства анализа неизвестных регрессионных зависимостей. Варьируя формой функции $w(x_i, x)$, можно получить различные версии алгоритмов сглаживания: скользящей средней, экспоненциальной моделью, ядерной функцией, сплайнами, локальной полиномиальной регрессией LOESS или LOWESS, методом Хольта-Винтерса и др. (Hastie at al., 2009).

Метод *локальной регрессии* (или LOESS – акроним от *lo*cal reg*ress*ion) представляет собой процедуру вычисления параметров линейной $y_t = \alpha_t + \beta_t x_t + \varepsilon_t$ или полиномиальной (квадратичной) $y_t = \alpha_t + \beta_t x_t + \gamma_t x_t^2 + \varepsilon_t$ модели с учетом вектора весовых коэффициентов, которые тем больше, чем ближе точки исходной выборки находятся по отношении к объекту x_t. Иными словами, для каждого текущего значения x_t коэффициенты α_t , β_t , γ_t подстраиваются динамически, выполняя скользящее усреднение данных в окне некоторой ширины. Важным параметром локальной регрессии является *степень сглаживания Span*, которая соответствует доле (фракции) от общего объема выборки, используемой для подбора коэффициентов в окрестностях точки t (Cleveland, 1979).

Аппроксимация *сплайнами* выполняет сглаживание в виде композиции из "кусочков" гладкой функции – как правило, полиномов. Свое происхождение термин "сплайн" ведет от длинных гибких линеек, используемых чертежниками в качестве лекал для проведения плавных кривых через заданные точки. Пусть отрезок оси абсцисс $[a, b]$ разбит на $(k + 1)$ частей точками $a_1 < ... < a_k$. Сплайном или кусочно-полиномиальной функцией степени m с k сопряжениями в точках a_1, ..., a_k называется функция $\varphi_j(x)$, которая на каждом интервале $(a_j, a_j + 1)$, $j = 0, ..., k$, описывается алгебраическим полиномом $P_j(x)$ степени m. Коэффициенты полиномов $P_j(x)$ согласованы между собой так, чтобы выполнялись условия непрерывности функции $\varphi_j(x)$ и ее $(m - 1)$ производных в узлах сопряжений. Поскольку функция имеет непрерывно дифференцируемые переходы в точках сочленения, участки состыковываются между собой так, что получившаяся кривая в целом также оказывается гладкой. Наиболее употребительны кубические сплайны с $m = 3$, которые обладает свойством минимальной кривизны (Розенберг и др., 1994).

В общем случае мерой близости некоторой кривой g к данным обучающей выборки является сумма квадратов невязок $MSE = \sum_{i=1}^{n} [y_i - g(x_i)]^2 / n$. Если не ограничивать количество узлов сплайна, то можно достичь хорошей аппроксимации данных, но получить кривую, имеющую очень много резких локальных изменений. Наоборот, плавность кривой достигается обычно за счет увеличения ошибки сглаживания. Компромисс между этими двумя противоречивыми целями достигается за счет введения

[7] Тут есть некоторая проблема с терминологией. При сглаживании, которое широко применяется в цифровой обработке, непрерывная функция не строится, и преобразуются только ординаты точек. Мы же рассматриваем варианты локальной аппроксимации для получения сглаженной функции. К сожалению, прямой перевод *smooth* fitted curve нам иного выхода не предоставляет.

штрафа за нарушение плавности, равного интегралу от квадрата второй производной $SPL = \int (g''(x))^2 dx$. Если ввести регулирующий параметр сглаживания λ, то функция оптимизации сплайна будет иметь вид $L(g, \lambda) = MSE + \lambda\ SPL$. Оптимальное значение λ находится с использованием методов ресамплинга (Maindonald, Braun, 2010).

Рассмотрим процедуры сглаживания на примере [П4] поиска зависимости массы внутренних органов животных от массы тела и других показателей. Для этого используем данные экспедиционных исследований, любезно предоставленные проф. А.В. Коросовым, по популяции красной полевки (*Clethrionomys rutilus*) в окрестностях г. Байкальска на расстоянии 5÷100 км от Байкальского целлюлозно-бумажного комбината (БЦБК). Для 278 особей, информация по которым была отобрана из базы данных, будем анализировать следующие морфометрические показатели:

- масса сердца (C), почек (R), надпочечников (Sr), печени (H), селезенки (L),

которые рассматривали как функции от других независимых переменных:

- массы тела (W), длины тела без хвоста (Lt), номера возрастной группы по классификации Тупиковой (age1) и месяца отлова (mon).

На рис. 4.8 показаны различные версии кривой, аппроксимирующей зависимость массы сердца от массы тела методом локальной квадратичной регрессии, при различных значениях степени сглаживания *Span* (по умолчанию используется *Span* = 0.75). 5% случайных членов исходной выборки использовались для экзамена и прогнозируемые значения массы сердца показаны на графике кружками с заполнением.

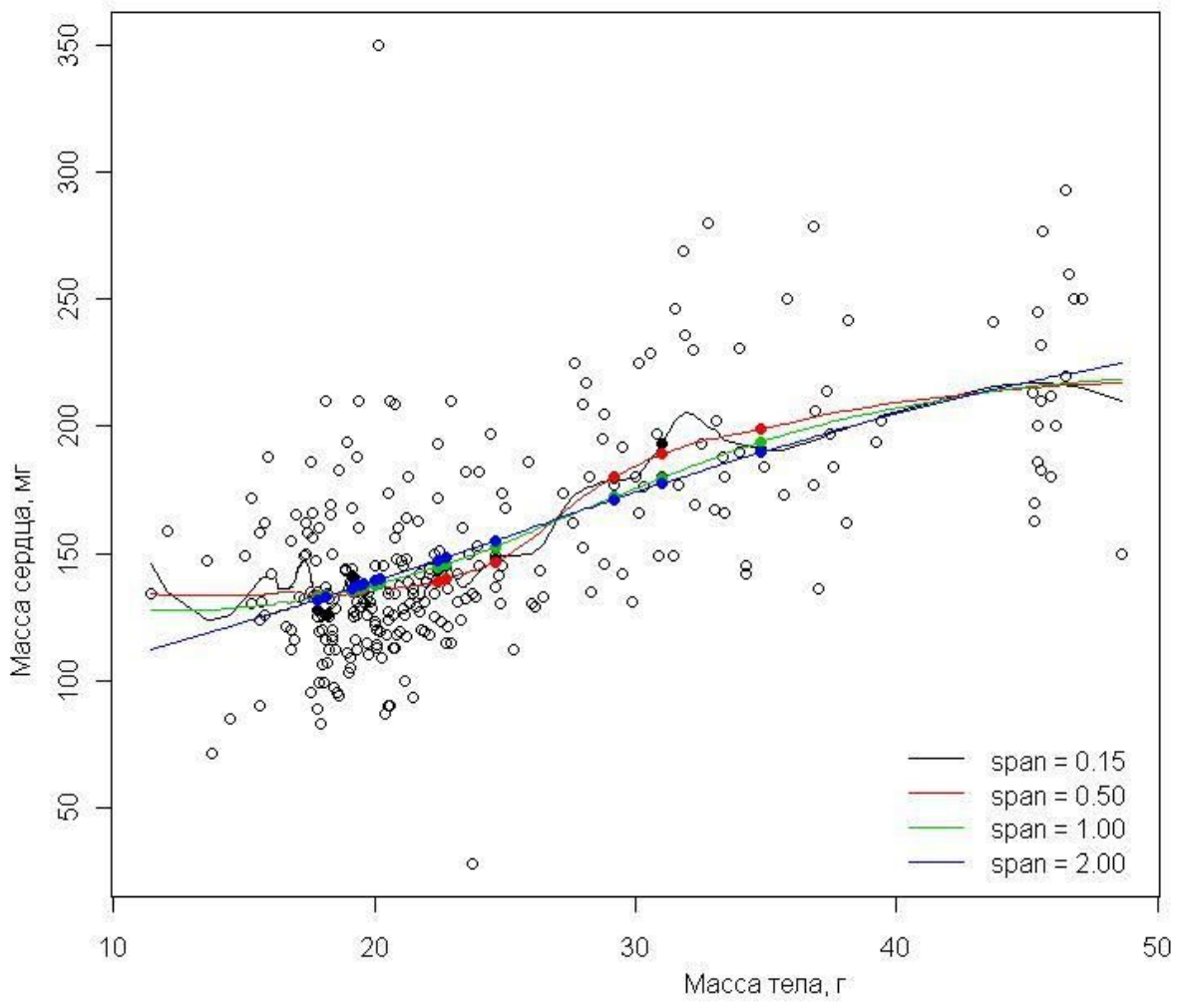


Рис. 4.8. Кривые сглаживания морфометрических зависимостей красной полевки методом локальной полиномиальной регрессии

На рис. 4.9 представлены варианты кубических сплайнов для аппроксимации зависимости массы печени от длины тела при различных значениях параметра сглаживания $\lambda = r \cdot 256^{(3Spar - 1)}$, где $r \leftarrow$ линейный функционал от матрицы наблюдений. При увеличении *Spar* плавность сглаживающей кривой увеличивается и при *Spar* = 1.5 она совпадает с прямой линейной регрессии. Оптимальный сплайн (*Spar* = 0.85, $\lambda = 0.0296$) можно найти с использованием процедуры скользящего контроля (или кросс-проверки leave-one-out – см. раздел 3.4).

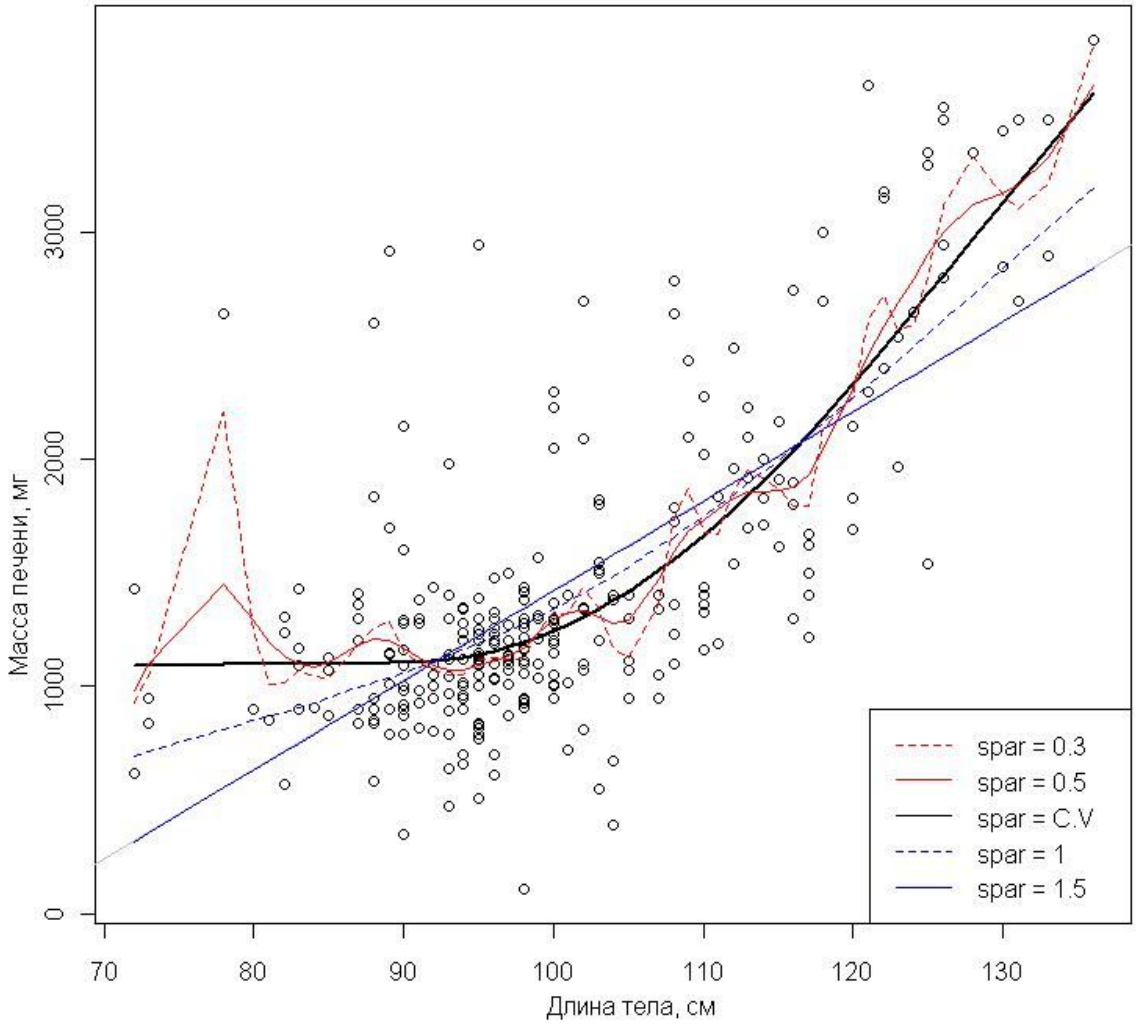


Рис. 4.9. Кубические сплайны для сглаживания морфометрических зависимостей красной полевки

В случае совместного статистического анализа небольшого числа факторов может быть проведено изучение характера их взаимной зависимости для всех возможных комбинаций пар (см. рис. 4.10). Однако, если число факторов достаточно велико и каждый из них оказывает специфическое влияние на значение зависимой переменной, то имеет смысл использовать методы многомерной регрессии.

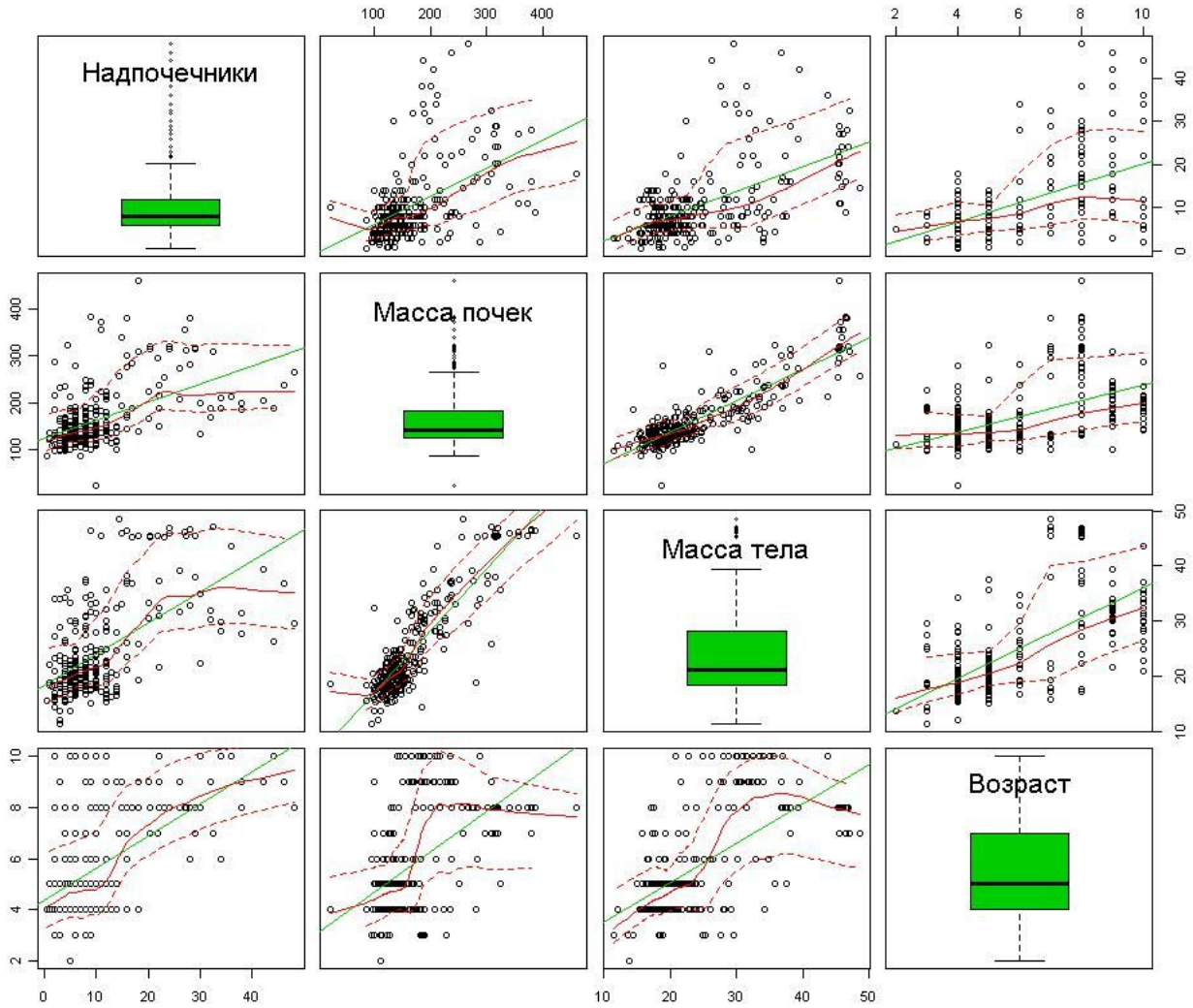


Рис. 4.10. Графики парных зависимостей четырех морфометрических показателей красной полевки; по диагонали – "боксы с усами" для отдельных переменных; в остальных клетках – линии линейной регрессии (зеленым цветом) и локальной регрессии с доверительными интервалами

Оценка значений отклика для регрессии от двух или более факторов в условиях существенной нелинейности изучаемых взаимодействий может быть выполнена с использованием обобщенных аддитивных моделей (GAM – Generalized Additive Models, см. Hastie, Tibshirani, 1990, Wood, 2006):

$$g(y) = \beta_0 + \sum_{i=1}^{p} f_i(x_i) + \varepsilon,$$

где $g(y)$ – функция связи, f_i - произвольные гладкие функции. В роли последних обычно используются обратная функция $f(x) = (1/x)$, логарифмическая функция $f(x) = \log(x)$, а также описанные выше непараметрические функции сглаживания локальной регрессией или сплайнами. В качестве вида функции связи мы ограничимся тождеством $g(y) = y$, но, например, в случае биномиального распределения широкое применение находят модели логита $f(y) = \log(y/(1 - y))$.

Аддитивные модели являются более гибкими инструментом аппроксимации, чем линейные модели, поскольку вид лучшей преобразующей функции для каждой переменной может быть тщательно подобран в ходе отдельных частных исследований. GAM представляют собой лишь средство обобщения частных функций в единое целое без каких-либо параметрических предположений относительно формы f_i.

Подбор компонентов наилучшей аддитивной модели имеет столь же важное значение, как и в случае обычной множественной регрессии. Одной из рекомендуемых

процедур (Burnham, Anderson, 2002) является ранжирование моделей по их информативности с использованием весов Акаике, которое заключается в следующем:

- строится коллекция моделей с перебором всех возможные сочетания факторов x и для каждой из них рассчитывается критерий Акаике $AIC = G^2 + 2k$, где G^2 – девианс (deviance), т.е. статистика, отражающая качество модели, которая оценивается как разность логарифмов оценок правдоподобия тестируемой модели и максимально насыщенной модели, k – число факторов, включенных в модель;
- построенные модели ранжируются по возрастанию AIC и вычисляется разность между критерием для каждой i-й модели и минимальным его значением $\Delta_i = AIC_i - AIC_{\min}$;
- подсчитываются веса Акаике $w_i = e^{-0.5\Delta_i} / \sum_i e^{-0.5\Delta_i}$ (см. табл. 4.6).

Таблица 4.6. Разности критерия Акаике Δ и веса w для выбора наилучшей аддитивной модели зависимости массы печени красной полевки (H) от возраста (age1), длины тела (Lt), месяца отлова (mon) и массы тела (W); lo(…) – функция сглаживания локальной линейной регрессией при *Span* = 0.5

№№	Коэффициенты модели					Статистики Акаике			
	β_0	lo(age1)	lo(Lt)	lo(mon)	lo(W)	df	AIC	Δ	w
11	253.6		-6.1		73.5	4	4066.	0	0.445
12	268.7	-12.2	-5.77		74.4	5	4067.1	1.057	0.262
15	105.8		-5.55	7.72	74.8	5	4067.7	1.685	0.191
16	169.1	-19.6	-5.01	1.45	76.6	6	4069.0	2.964	0.101
10	-158.9	-26.7			71.6	4	4084.8	18.7	0.00004
9	-231.1				68.3	3	4086.1	20.0	0.00002
14	-283.8	-29.4		9.81	74.2	5	4087.0	21.053	0.00001
13	-429.6			17.9	70.7	4	4088.1	22.0	0
3	-2530		39.5			3	4173.8	107.8	0
8	-2599	37.2	35.9	27.1		5	4174.7	108.6	0
7	-2786.5		40.9	14.0		4	4175.1	109.0	0
4	-2321	23.5	36.1			4	4175.4	109.4	0
6	-150.3	189.6		63.1		4	4282.4	216.3	0
2	465.0	168.3				3	4302.3	236.3	0
5	1837.8			-52.5		3	4405.6	339.6	0
1	1427.2					2	4421.8	355.8	0

Представленные веса w_i интерпретируются (Burnham, Anderson, 2002) как вероятности того, что i-я модель является лучше, чем любая другая на множестве моделей-претендентов. Считается, что если веса отличаются менее, чем на 10% от $w_{\max}$, то эти модели идентичны по качеству наилучшей. Так в табл. 4.6 модель 11 в 2.5 раза "лучше" модели 15 с добавлением еще одного фактора (mon) и в 4.5 раза "лучше" полной модели 16. С другой стороны, сравнение остаточных дисперсий для моделей 11 и 16 не дает статистически значимых отличий по критерию Фишера ($p = 0.11$).

С использованием девианса G^2 аддитивной модели можно рассчитать соотношение между суммой отклонений для остатков *RD* (Residual Deviance) и общей суммой отклонений *ND* (Null Deviance). Тогда статистика $R^2 = (1 - RD/ND)$ может быть интерпретирована как аналог коэффициента детерминации, который обычно используется для оценки линейной регрессии. Для модели 11 из табл. 4.6, оценивающей зависимость массы печени, эта статистика будет равна $R^2 = 0.738$. Для сравнения простая линейная модель с параметрами будет иметь вид:

$$H = 664.8 + 85W - 13\,Lt,$$

при статистически значимых оценках коэффициентов β_1 , β_2, но несколько меньшем коэффициенте детерминации $R^2 = 0.703$.

Можно также отметить, что аналогичная функция gam() в другом пакете *mgcv* реализует модель со смешанными эффектами, что приводит к автоматическому выбору степени сглаживания на основе обобщенной кросс-проверки.

В результате построения аддитивных моделей нами получено для каждого показателя, связанного с массой внутренних органов красной полевки, по три вектора значений: (а) натуральные измерения $\boldsymbol{x}$, (б) рассчитанные по моделям GAM $\boldsymbol{x}_{\text{fit}}$, изменчивость которых определяется только массо-размерными параметрами животного, а влияние всех остальных факторов "снято", (в) остатки $\boldsymbol{x}_{\text{res}} = \boldsymbol{x} - \boldsymbol{x}_{\text{fit}}$, объединяющие погрешность измерений и вариацию массы внутренних органов под воздействием любых иных факторов, кроме четырех предикторов модели. Интересно сравнить результаты дисперсионного анализа этих переменных в зависимости от таких факторов как "пол" (самцы / самки) и местообитание M (1 – в зоне до 5 км от БЦБК, 2 – 15 ÷ 50 км, 3 – свыше 50 км) – см. табл. 4.7. Из расчетов можно предположить, что статистически значимая связь массы почек мыши с местообитанием определяется не патологией внутренних органов, а только тем, что за пределами зоны влияния комбината отлавливаемые особи оказывались крупнее. Вполне понятно также, что остаточная изменчивость показателя определяется в основном гендерными отличиями.

Таблица 4.7. Двухфакторный дисперсионный анализ показателей массы почек красной полевки в зависимости от пола и местообитания M (*df* – число степеней свободы, *MSS* – средние суммы квадратов, *F* – статистика Фишера, Pr(>*F*) – *p*-значение)

Фактор	*df*	Измеренные $\boldsymbol{x}$			Модельные $\boldsymbol{x}_{\text{fit}}$			Остаточные $\boldsymbol{x}_{\text{res}}$		
		MSS	*F*	Pr(>*F*)	*MSS*	*F*	Pr(>*F*)	*MSS*	*F*	Pr(>*F*)
M	1	37922	9.38	**0.0024**	38824	12.58	**0.0004**	5.3	0.0062	0.937
Пол	1	4143	1.02	0.312	20576	6.67	**0.0103**	6252	7.30	**0.007**
Остатки	275	4039			3085			856		

К разделу 4.7:

```
#  Загрузка данных из файла Excel и их подготовка
library(xlsReadWrite) ; library (gam); library(car); library(MuMIn)
TOR <- read.xls("Ruts.xls", sheet = 1, rowNames=TRUE) ; attach (TOR)
# Создаем отсортированные векторы и список значений параметра сглаживания
x<-W[order(W)] ; y <-C[order(W)] ; spanlist <- c(0.15, 0.50, 1.00, 2.00)
# 15 случайных пар значений выделим для экзамена
MissingList <- sample(x,15); x[MissingList] <- NA
# Рисуем облако точек обучающей выборки
plot(x,y, xlab="Масса тела, г", ylab="Масса сердца, мг")
# Выводим кривые сглаживания LOESS для различных значений параметра span
for (i in 1:length(spanlist)) {
  y.loess <- loess(y ~ x, span=spanlist[i], data.frame(x=x, y=y))
  y.predict <- predict(y.loess, data.frame(x=x)) ;  lines(x,y.predict,col=i)
  # Рисуем точками прогнозируемые значения для экзаменационной последовательности
  y.Missing <-  predict(y.loess, data.frame(x=MissingList))
  points(MissingList, y.Missing, pch=FILLED.CIRCLE<-19, col=i) }
legend ("bottomright", c(paste("span =", formatC(spanlist, digits=2, format="f"))),
        lty=SOLID<-1, col=1:length(spanlist), bty="n")
# Выводим кривые сглаживания сплайнами для различных значений параметра spar
x<-Lt[order(Lt)] ; y <-H[order(Lt)]
plot(x,y,xlab="Длина тела, см", ylab="Масса печени, мг") ; abline(lm(y ~ x),col="grey")
sp.spline <- smooth.spline(x,y,cv=TRUE) ; lines(sp.spline, lwd=2) # Оптимальная кривая
lines(smooth.spline(x,y,spar=1.5),col="blue");
lines(smooth.spline(x,y,spar=1),col="blue",lty=2)
lines(smooth.spline(x,y,spar=0.5),col="red")                                    ;
lines(smooth.spline(x,y,spar=0.3),col="red",lty=2)
legend ("bottomright", c("spar = 0.3",  "spar = 0.5", "spar = C.V", "spar = 1","spar = 1.5"),
                    col = c("red", "red", 1, "blue", "blue"), lty = c(2,1,1,2,1))
```

```
# Выводим  матрицу графиков рассеивания для четырех показателей
Labels <- c("Надпочечники","Масса почек", "Масса тела", "Возраст")
scatterplotMatrix(~SR + R +  W + age1,var.labels = Labels, data = TOR, diag = "boxplot")
# Подбор моделей GAM
dredge(gam(C ~ lo(W) + lo(Lt) + lo(age1) + lo(mon),
                            family = gaussian, data = TOR), rank = "AIC")
dredge(gam(R ~ lo(W) + lo(Lt) + lo(age1) + lo(mon),
                            family = gaussian, data = TOR), rank = "AIC")
dredge(gam(SR ~ lo(W) + lo(Lt) + lo(age1) + lo(mon),
                            family = gaussian, data = TOR), rank = "AIC")
dredge(gam(H ~ lo(W) + lo(Lt) + lo(age1) + lo(mon),
                            family = gaussian, data = TOR), rank = "AIC")
dredge(gam(L ~ lo(W) + lo(Lt) + lo(age1) + lo(mon),
family = gaussian, data = TOR), rank = "AIC")
#  Расчет оптимальных моделей  GAM и вывод модельных значений и остатков
C.gam <- gam(C ~ lo(W) + lo(Lt) + lo(age1) + lo(mon), family = gaussian, data = TOR)
R_1.gam <- gam(R ~ lo(W) + lo(Lt) , family = gaussian, data = TOR)
SR_1.gam <- gam(SR ~ lo(age1)+ lo(Lt)+ lo(mon) , family = gaussian, data = TOR)
H_1.gam <- gam(H ~ lo(W) + lo(Lt) , family = gaussian, data = TOR)
L_1.gam <- gam(L ~ lo(W) + lo(Lt) , family = gaussian, data = TOR)
TOR.fit <- data.frame(TOR, C.res = C.gam$res , C.fit = C.gam$fit,
       R.res=R_1.gam$res,R.fit=R_1.gam$fit,SR.res=SR_1.gam$res,SR.fit=SR_1.gam$fit,
       H.res=H_1.gam$res,H.fit=H_1.gam$fit,L.res=L_1.gam$res,L.fit=L_1.gam$fit)
save(TOR.fit, file="Ruts.RData")  #  Сохраняем расширенную таблицу для дальнейших расчетов
H.gam <- gam(H ~ lo(W) + lo(Lt) + lo(age1) + lo(mon), family = gaussian, data = TOR)
summary(H.gam) ; anova(H.gam, H_1.gam)  #  Сравниваем оптимальную и полную модели
plot(H_1.gam,residuals=TRUE,se=TRUE, ask =TRUE)  #  Выводим графики остатков
summary(lm(H ~ W + Lt , family = gaussian, data = TOR)) #  линейная модель для сравнения
#  Двухфакторный  дисперсионный анализ зависимости массы почек от пола и местообитания
attach(TOR.fit) ; summary(aov(R ~ M + sex))
summary(aov(R.fit ~ M + sex)) ; summary(aov(R.res ~ M + sex))
```

4.8. Многомерный анализ MANOVA и метод случайного зондирования

В предыдущих разделах мы рассматривали методы, позволяющие выяснить, как влияет на значение одномерного отклика **Y** совокупность *m* факторов **X**, выраженных независимыми непрерывными переменными или категориями группирующих признаков. Предположим теперь, что для каждого из *n* изучаемых объектов измерено *p* переменных, которые мы считаем откликом, и необходимо оценить эффекты влияния факторов не на одну, а на весь комплекс из *p* признаков в совокупности, т.е. на многомерную зависимую переменную (Seber, 2004).

Рассмотрим предварительно задачу сравнения векторов средних $\overline{\mathbf{X}}_1, \overline{\mathbf{X}}_2$ для двух популяций, заданных подматрицами наблюдений $\mathbf{X}_1$ размерностью $n_1 \times m$ и $\mathbf{X}_2$ ($n_1 \times m$). В разделе 2.5 мы в аналогичном случае проверяли последовательно четыре нулевых гипотезы, вычисляя *t*-критерии и *p*-значения для каждой переменной, и использовали метод Бонферрони, выполняющей множественные сравнения. Альтернативный многомерный подход оперирует единственной тестовой статистической величиной, которая принимает во внимание различия выборочных средних для всех переменных в совокупности.

При многомерном анализе для проверки статистических гипотез в целом используются те же статистические критерии, что и в одномерном, однако они модифицированы с учетом природы многомерных случайных величин. Как и в одномерном случае оценки равенства средних по *t*-критерию (раздел 2.3), предполагается, что в совокупности имеет место нормальное распределение $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ координат точек объектов x_{ij} в многомерном пространстве. Тогда для *m* измеренных переменных мы имеем вектор средних $\boldsymbol{\mu}$ размерностью *m* и матрицу ковариаций $\boldsymbol{\Sigma}$ ($m \times m$). При разделении исходной выборки *n* на 2 класса ($n_1 + n_2$) сдвиг относительно друг друга центроидов

многомерного распределения точек может быть рассчитан как обобщенное расстояние Махалонобиса $$D^2 = (\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2)^t \mathbf{C}_w^{-1} (\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2),$$

где $\overline{\mathbf{X}}_1, \overline{\mathbf{X}}_2$ – векторы оценок групповых средних значений координат, $\mathbf{C}_w$ – объединенная матрица оценок внутригрупповых ковариаций. Обычно используют простейшую формулу для объединения выборочной ковариационной матрицы: $\mathbf{C}_w = \dfrac{(n_1 - 1)\mathbf{C}_1 + (n_2 - 1)\mathbf{C}_2}{n_1 + n_2 - 2}$,

где $\mathbf{C}_1$ и $\mathbf{C}_2$ – стандартные оценки ковариационных матриц по каждой из двух выборок.

Статистическая проверка гипотезы о равенстве двух векторов $H_o: \overline{\mathbf{X}}_1 = \overline{\mathbf{X}}_2$ может быть осуществлена с использованием двухвыборочной статистики Хотеллинга, которая является многомерным аналогом *t*-критерия Стьюдента:

$$T^2 = \frac{n_1 n_2}{n_1 + n_2} (\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2)^t \mathbf{C}_w^{-1} (\overline{\mathbf{X}}_1 - \overline{\mathbf{X}}_2) = \frac{n_1 n_2}{n_1 + n_2} D^2.$$

При справедливости H_o величина $F = \dfrac{n_1 + n_2 - m - 1}{(n_1 + n_2 - 2)m} \cdot T^2$ имеет *F*-распределение с m и $(n_1 + n_2 - m - 1)$ степенями свободы.

Рассмотрим в качестве примера [П1] пространственно-временную изменчивость зоопланктонных сообществ Куйбышевского водохранилища, выявленную в ходе многолетних наблюдений. Отклик экосистемы **Y** будем оценивать по 4 переменным, соответствующим биомассе отдельных таксономических групп организмов: каляноид (CAL), кладоцер (CLA), циклопоид (CYC) и коловраток (ROT). Выполним сравнение двух выборок наблюдений, сделанных в разных точках акватории водохранилища до 1966 г. (195 строк исходной таблицы) и в последующий период (195 строк). Расчет по приведенным формулам привел к следующим результатам: статистика Хотеллинга $T^2 = 97.4$, $F = 24.2$, $p \cong 0$. Статистически значимые структурные изменения зоопланктоценоза, которые произошли через 8 лет после ввода в строй плотины Волжской ГЭС, выразились в увеличении биомассы кладоцер и коловраток, в то время как популяционная плотность каляноид и циклопоид существенно уменьшилась.

Другой способ оценить *p*-значение – это выполнить рандомизацию, для чего достаточно провести многократное перемешивание между собой строк сравниваемых выборок и построить статистическое распределение величины статистики Хоттелинга T^2 при справедливости нулевой гипотезы. Доля случаев, когда T^2_{ran} для псевдо-выборок превысил бы соответствующее эмпирическое значение $T^2_{\text{obs}} = 97.4$, даст нам оценку вероятности ошибки 1-го рода. В нашем примере после 1000 итераций таких случаев не оказалось, т.е. $p = 0.001$. Заметим, что использование любой из статистик, расстояния Махалонобиса D^2, T^2 или *F*-критерия, которые связаны между собой только постоянным множителем, всегда приведет к одному и тому же результату. Таким образом, современная доступность точных (или "почти точных") рандомизационных методов практически снимает необходимость контролировать сходимость используемого тестового критерия T^2 к известному распределению статистики Фишера. Более того, сами эти статистики становятся "ненужной частностью Оккама", уступив место предметно интерпретируемому расстоянию Махалонобиса. Детальный обзор рандомизационных тестов, основанных на дистанциях, приведен в работе (Mielke, Berry, 2001).

Но не является ли временная изменчивость зоопланктонных сообществ следствием пространственной неоднородности? Разделим акваторию Куйбышевского водохранилища на 5 участков по отдельным плесам: (1) Волжский от Чебоксарской ГЭС до устья Камы, (2) Волго-Камский, (3) Тетюшинский + Ундорский, (4) Ульяновский и (5) Новодевиченский + Приплотинный. Поставим задачу оценки статистической значимости такого разбиения в условиях временной динамики.

Многомерный дисперсионный анализ MANOVA позволяет проверить не только гипотезы о влиянии набора m независимых факторов $\mathbf{X}$ на каждую из q зависимых переменных Y_i в отдельности, но и гипотезу о влиянии факторов на всю совокупность многомерного отклика $\mathbf{Y}$. Современная процедура MANOVA основана на многомерной обобщенной линейной модели (Афифи, Эйзен, 1982) $\mathbf{Y} = \mathbf{X}'\boldsymbol{\beta} + \boldsymbol{\varepsilon}$,
где $\boldsymbol{\beta}$ ($m \times p$) –матрица неизвестных параметров, а $\boldsymbol{\varepsilon}$ ($n \times p$) – матрица остатков, строки которой составляют случайную выборку размера n из невырожденного q-мерного распределения $N(\mathbf{0}, \boldsymbol{\Sigma})$. Такой подход предоставляет исследователю целый ряд важных дополнительных возможностей анализа.

Если векторы коэффициентов модели оценены, например, методом МНК, то можно сформировать общую матрицу сумм квадратов отклонений и парных произведений векторов **T,** из которой можно вычленить остаточную (внутригрупповую) сумму квадратов **W**. Третью матрицу **B** = **T** - **W** называют матрицей сумм квадратов и произведений, *обусловленных отклонением от нулевой гипотезы* (cross-product). В работе Рао (Rao, 1965) показано, что для проверки H_0 требуется определить q корней $\lambda_1, \lambda_2, \ldots, \lambda_q$ характеристического уравнения $|\mathbf{W} - \lambda \mathbf{T}| = 0$, поэтому для тестирования могут быть использованы различные функции, зависящие от λ_q.

Наиболее часто при расчетах в качестве тестовой статистики используется Λ-критерий Уилкса (Wilks): $\Lambda = |\mathbf{W}| / |\mathbf{T}| = |\mathbf{W}| / |\mathbf{B} + \mathbf{W}| = \lambda_1\lambda_2\ldots\lambda_q$, который изменяется от 0 (группы имеют статистически значимые отличия) до 1 (нет влияния группировочных факторов). Кроме критерия Λ Уилкса разработан целый арсенал различных статистик, которые тем или иным способом учитывают характер многомерного распределения ковариаций и выводятся в статистических программах под наименованиями:

- "Pillai's Trace" (след Пиллая) $V = \sum_{ij}(\mathbf{B}+\mathbf{W})^{-1}\mathbf{B} = \sum_i \lambda_i/(1+\lambda_i)$;
- "Hotelling-Lawley's Trace" (след Хотеллинга-Лавли) $\tau = \sum_{ij}\mathbf{W}^{-1}\mathbf{B} = \sum_i \lambda_i$;
- "Roy's Largest Root" (максимальный характеристический корень Роя) $\theta = (\mathbf{B}+\mathbf{W})^{-1}\mathbf{B} = \lambda_1/(1+\lambda_1)$.

С помощью разработанных аппроксимирующих зависимостей по этим тестовым величинам восстанавливается распределение значения F и оценивается статистическая значимость p проверяемой гипотезы.

В частном случае при k разбиениях одной независимой переменной многомерный дисперсионный анализ проверяет нулевую гипотезу H_0: $\mu_1 = \mu_2 = \ldots = \mu_k$, где $\boldsymbol{\mu}$ – вектор средних для уровней влияющего фактора. При числе независимых переменных $m > 1$ оценивается статистическая значимость коэффициентов $\boldsymbol{\beta}$ обобщенной линейной модели. Для рассматриваемого примера была выявлена достоверность влияния обоих факторов, причем оценки значимости по всем тестовым статистикам оказались чрезвычайно близки (в пределах четырех значащих цифр аппроксимируемой статистики Фишера). Значимое парное произведение факторов (табл. 4.8) свидетельствует о неоднородности многолетних изменений структуры зоопланктона на различных участках водохранилища.

Таблица 4.8. Двухфакторный многомерный дисперсионный анализ распределения групп зоопланктона по участкам Куйбышевского водохранилища в периоды до 1966 г. и после (F – статистика Фишера, Pr(>F) – p-значение)

Компоненты двух-факторной модели	Pillai's Trace V	Wilks' Lambda Λ	Hotelling-Lawley's τ	Roy's Root θ	F (аппрокс.)	Pr(>F)
Зоны водохрани-лища (Zone)	0.174	0.825	0.210	0.210	20.2	< 0.0001
Период наблюде-ний (F)	0.203	0.796	0.254	0.254	24.4	< 0.0001
Zone : F	0.043	0.956	0.045	0.045	4.37	0.0018

Рассмотрим теперь метод *случайного зондирования*, реализующий другой интересный (и незаслуженно забытый) подход к оценке изменчивости экосистемных компонент, не связанный напрямую с многомерной статистикой и разработанный учеными-экологами.

Пусть мы имеем матрицу **X**, включающую наблюдения *m* показателей для *n* объектов. Каждое *i*-е измерение из *n* выполняется в различных условиях: в разные моменты времени, точках пространства или при монотонном изменении любого иного воздействия. Как уже обсуждалось в разделе 3.6, упорядоченную последовательность **Y** значений произвольного независимого фактора в экологии принято называть *градиентом*: широтный градиент, высотный градиент, температурный градиент, продольный градиент реки и т.д.

Необходимо проверить нулевую гипотезу, что данные многомерных наблюдений в своей совокупности не зависят от изучаемого градиента **Y**. Обычно "интересные" линейные комбинации таких зависимостей в многомерных наборах данных выделяются с использованием различных методов оптимального целенаправленного проецирования (projecting pursuit – Зиновьев, 2000). Для этого проводится редукция матрицы **X** в пространства малой размерности (с 2 или 3 осями координат), что обеспечивает наглядное графическое представление исследуемых объектов. Однако целенаправленное проецирование вовлекает в процесс оптимизации геометрической метафоры различные предположения и индексы сходства, а также не дает прямого ответа о справедливости нулевой гипотезы.

Э. Пиелу (Pielou, 1984) разработала непараметрический метод случайного зондирования (random skewers), который позволяет объективно судить, насколько статистически значима детерминированная тенденция в изменении структуры всего набора данных **X** в целом вдоль изучаемого градиента. Рассмотрим простейший случай, когда точки измерений были выстроены в определенном порядке 1, 2, ..., *n* согласно их положению во времени или пространстве (например, по течению водотока), а последовательность **Y** представляет собой натуральный ряд чисел от 1 до *n*. Обобщение на произвольный характер значений градиента рассматривается Б. Манли (Manly, 2007).

Данные **X** представляются как "облако" (swarm) *n* точек, соответствующих выполненным измерениям, в *m*-мерном пространстве исходных признаков. "Зонд", пронизывающий эту *n*×*m*-мерную структуру, представляет собой произвольно ориентированный вектор, координаты которого определяются *n* случайными числами, нормированными так, чтобы их сумма равнялась 1. Из каждой точки "облака" данных на зонд опускается перпендикуляр и фиксируется место его пересечения со случайным вектором. Далее вычисляется коэффициент ранговой корреляции τ Кендалла между порядком следования проекций *n* точек на зонд и значениями градиента **Y**. Если величина τ близка к 1 (или -1), то делается вывод, что результаты наблюдений имеют очевидную закономерную упорядоченность относительно некоторого направления в многомерном пространстве. При $\tau \cong 0$ можно предположить одну из двух причин:

° все экспериментальные точки беспорядочно расположены по всем направлениям или

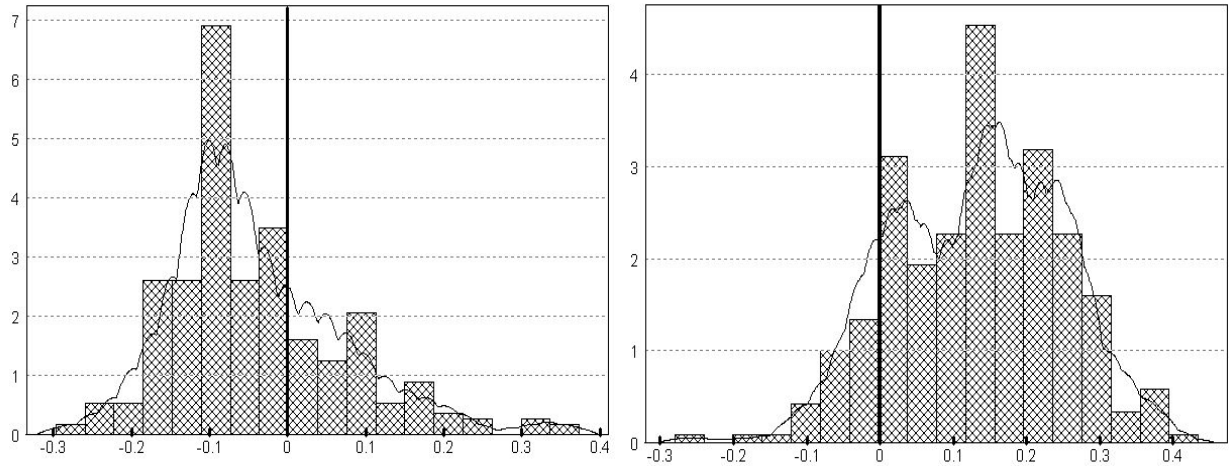
° ориентация случайного зонда не совпадает ни с одной из возможных осей многомерного эллипсоида "облака" точек эмпирических данных.

Процедура "пронизывания данных зондом" проводится многократно (например, 300 раз), и формируется частотное распределение всех зарегистрированных значений коэффициента ранговой корреляции τ. Если в результате получается унимодальное распределение с математическим ожиданием $\tau \cong 0$, то в "облаке" точек нет ни одного направления, относительно которого эмпирические данные упорядочены вдоль изучаемого градиента, что соответствует нулевой гипотезе. Согласно альтернативной гипотезе, принимаемой, когда центр распределения $|\tau_{sim}| > 0$, точки наблюдений закономерно упорядочены относительно оси фактора. Бимодальное распределение с

центрами, приближенными к +1 и -1, свидетельствует о существовании ярко выраженной детерминированной закономерности. Несколько вариантов проверки нулевой гипотезы об унимодальности эмпирического распределения τ и значимости его сдвига относительно нуля, представлено в работах (Pielou, 1984; Perry, Schaeffer, 1987).

Выполним тест случайного зондирования Пиелу с использованием данных о частоте встречаемости m = 129 различных видов макрозообентоса на каждой из n = 13 станций наблюдения р. Сок [пример П2], которые пронумеруем от 1 (исток) до 13 (устье). Для этой композиции видов получена гистограмма эмпирического распределения значений τ (рис. 4.11а), которая имеет достаточно выраженный унимодальный характер. При этом доверительный интервал для центра распределения не включает нулевое значение, что позволяет отклонить нулевую гипотезу. В качестве основного итога можно предположить в целом умеренную изменчивость тестируемого сообщества вдоль продольного градиента водотока.

В другом варианте расчета для тех же станций р.Сок ограничим список видов, выбрав только 66 таксонов, относящихся к одному семейству Chironomidae. Можно отметить (рис. 4.11б), что сдвиг относительно нуля центра распределения τ для ценоза хирономид носит более акцентированный (и противоположный по знаку!) характер, чем для всего бентоценоза. Д. Перри и Д. Шеффер (Perry, Schaeffer, 1987) выполняли тест случайного зондирования на различных локальных группах бентосных организмов, комбинируя их по таксономической и функциональной принадлежности, и добивались при этом отчетливого "расщепления" пиков распределения значений τ.



а) для 129 видов макрозообентоса (правая доверительная граница среднего $m + t \cdot s_m$ = -0,04 + 1,96·0,0065 = -0,027)

б) для 66 таксонов Chironomidae (левая доверительная граница среднего $m - t \cdot s_m$ = 0,13 – 1,96·0,0067 = 0,119)

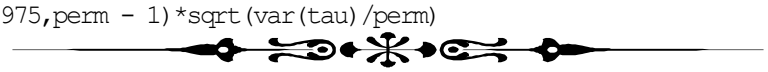
Рис. 4.11. Гистограмма распределения коэффициента ранговой корреляции Кендалла, полученная в ходе 300 итераций по методу случайного зондирования

Это вполне объяснимо, поскольку разные группы организмов имеют различный характер адаптационных реакций на изменяющиеся биотопические условия (а также различные стратегии жизненных циклов организмов). Чем разнообразнее состав сообщества, тем слабее выражена его совокупная экосистемная зависимость от градиента фактора. В этой связи каждый раз надо отчетливо представлять смысл той задачи, которая стоит перед исследователем, и какая нулевая гипотеза нуждается в проверке.

К разделу 4.8:

```
#  Загрузка данных из файла Excel и их подготовка
library(xlsReadWrite) ; PT <- read.xls("Куйб.xls", sheet = 1, rowNames=FALSE)
F <- as.factor(findInterval(PT[,4],c(0,1966))) ; Y <- as.data.frame(PT[,5:8])
#   Многомерный анализ
```

```
#   Сравнение  двух групп многомерных данных по критерию Хотеллинга
hotelling = function(y, f) {    #   Функция обработки: y – общая матрица, f – фактор
группировки
  k = ncol(y); y1 <- y[which(f==1),] ; y2 <- y[which(f==2),] ; n1 = nrow(y1); n2 = nrow(y2)
  ybar1= apply(y1, 2, mean); ybar2= apply(y2, 2, mean); diffbar = ybar1-ybar2
  v = ((n1-1)*var(y1)+ (n2-1)*var(y2)) /(n1+n2-2)       #   общая ковариационная матрица
  t2 = n1*n2*diffbar%*%solve(v)%*%diffbar/(n1+n2)   #   статистика Хотеллинга
       # критерий Фишера и p-значение
  f = (n1+n2-k-1)*t2/((n1+n2-2)*k) ;  pvalue = 1-pf(f, k, n1+n2-k-1)
  return(list(pvalue=pvalue, f=f, t2=t2, diff=diffbar))   }
hotelling(Y, F)    #    Расчет по созданной функции обработки данных
source("print_rezult.r") ; Nrand = 1000 ; reject = numeric(Nrand)   #   Рандомизационный тест
#   Формируем распределение F-критерия
for (i in 1:Nrand) {reject[i] = hotelling(Y, sample(F))$f}
RandRes(as.numeric(hotelling(Y, F)$f) , reject, Nrand)
library(Hotelling)  #    То же самое, но с использованием функций пакета Hotelling
split.data <- split(Y,F)  ;  summ.hots <- hotelling.stat(split.data[[1]], split.data[[2]])
summ.hot <- hotelling.test(split.data[[1]], split.data[[2]], perm = T, B = 1000)
#   Выполнение многомерного дисперсионного анализа по двум факторам: зонам и периоду
model11<-manova(cbind(BCAL, BCLA, BCYC,BROT)~Zone, data=PT)
model12<-manova(cbind(BCAL, BCLA, BCYC,BROT)~F, data=PT)
model<-manova(cbind(BCAL, BCLA, BCYC,BROT)~Zone*F, data=PT)
anova(model) ; summary(model,test="W"); summary(model,test="H") ; summary(model,test="R")
#    ------------------------------------------------------------------------
#    Метод случайного зондирования
A <- read.xls("Сок1.xls", sheet = 1, rowNames=TRUE)
A[is.na(A)] <- 0 ; n <-ncol(A) ; m <-nrow(A)
#  Задание вектора с градиентом и числа итераций
Y <-1:n ; perm <- 300 ; tau <- numeric(perm)
#  Определение функции, вычисляющей расстояние от начала координат в направлении
#  случайного зонда Skewer  до  пересечения с перпендикуляром, опущенным на Skewer
#  из произвольной точки Ajv в  m-мерном пространстве
Rast <- function (Ajv) {
       XZ <- numeric(m) ; XZ[1] =  Skewer[1] * sum(Ajv*Skewer) / sq
       for (i in 2:m) XZ[i] <- XZ[1] * Skewer[i] / Skewer[1]
       return (sqrt(sum(XZ*XZ))) }
#  Выполнение итераций вычисления распределения коэффициентов ранговой корреляции
for (ipa in 1:length(tau)) {
       # создание случайного зонда
       Skewer <- runif(m, max=m); sq <- sum(Skewer*Skewer)
       #  вычисление вектора проекций эмпирических точек на случайный зонд
       X <- sapply(1:n, function (j) { Rast(as.vector(A[,j]))})
       # вычисление коэффициента тау Кендалла
       tau[ipa] <- (cor(X,Y, method = "kendall")) }
# вывод результатов имитации (гистограммы и доверительных интервалов на основе t-критерия)
hist(tau) ; mean(tau) ; tau ;
mean(tau) - qt(0.975,perm - 1)*sqrt(var(tau)/perm)
mean(tau) + qt(0.975,perm - 1)*sqrt(var(tau)/perm)
```



5. МЕТОДЫ, ИСПОЛЬЗУЮЩИЕ МАТРИЦЫ ДИСТАНЦИЙ

5.1. Меры сходства/расстояния в многомерном пространстве

Наиболее содержательные разделы статистических исследований связаны с многомерным анализом данных, когда каждый элемент изучаемой системы описывается множеством переменных. При этом важно постулировать тип модели информативного пространства, чтобы в рамках поставленной задачи наиболее корректно задать количественную меру отношений между объектами.

В большинстве случаев многомерное пространство понимается как множество измеренных переменных: фиксированных или случайно варьируемых факторов (в экологии, например, условия среды или наличие ресурсов), которые потенциально определяют наблюдаемые свойства исследуемых объектов (обилие видов, показатели здоровья и т.д.). С формальных позиций это не пространство, а предметно-ориентированная система мониторинга, структура информативного пространства которой строго не определена. В частном случае при применении статистических методов пространство измерений может интерпретироваться как вероятностное: тогда пара векторов действительных чисел $\{x_1, x_2,\ldots, x_m\}$ и $\{y_1, y_2,\ldots, y_m\}$, описывающих произвольные объекты x и y, будут трактоваться как выборочные реализации m-мерной случайной величины. В определенном смысле в качестве мер сходства между этими объектами могут выступать оценки ковариации $\mathrm{cov}(x, y) = \sum_{i=1}^{m}(x_i - m_x)(y_i - m_y)$, коэффициент корреляции $r_{xy} = \mathrm{cov}(x, y)/\sigma_x\sigma_y$ или произвольное ковариационное отношение $K = [\mathrm{cov}(x, y) - \mathrm{cov}_{\min}]/(\mathrm{cov}_{\max} - \mathrm{cov}_{\min})$, где m – математическое ожидание, σ – стандартное отклонение, $\mathrm{cov}_{\min}$ и $\mathrm{cov}_{\max}$ – экстремальные значения ковариации для теоретической ("эталонной") выборки (Воробейчик, 1993).

В общем случае использование вероятностного пространства совершенно не обязательно. Часто пространство измеряемых переменных рассматривают как *метрическое* пространство, расстояния в котором определяются некоторой функцией ρ, обладающей нехитрыми свойствами: а) тождества $\rho(\mathbf{x}, \mathbf{y}) = 0$ при $\mathbf{x} = \mathbf{y}$, б) симметрии $\rho(\mathbf{x}, \mathbf{y}) = \rho(\mathbf{y}, \mathbf{x})$ и в) правила треугольника $\rho(\mathbf{x}, \mathbf{y}) + \rho(\mathbf{y}, \mathbf{z}) \geq \rho(\mathbf{x}, \mathbf{z})$. Конкретной дефиницией функции ρ может быть, например, обобщенная мера Минковского, наиболее популярными вариантами которой являются манхеттенское и евклидово расстояния.

В экологии меру сходства видового состава двух сообществ **x** и **y** часто рассматривают как долю числа совпадающих особей относительно средней численности особей всех видов в этих сообществах и оценивают с использованием количественной меры Брея–Кёртиса[8]: $M_{xy} = 2\sum_{i=1}^{m_{ab}}\min[x_i, y_i]/(\sum_{i=1}^{m}x_i + \sum_{i=1}^{m}y_i)$. Нетрудно заметить, что величина $d_{xy} = (1 - M_{xy})$ соответствует нормированному манхеттенскому расстоянию. Можно отметить другие меры сходства, также принимающие значения от 0 до 1, но нормировка которых имеет более сложный характер, компенсирующий негативные статистические эффекты – см. рекомендации в статье (Boyce, Ellison, 2001):

индекс Мориситы-Хорна $M_{xy} = \dfrac{\sum_{i=1}^{m}[(x_i + y_i)\log(x_i + y_i)] - \sum_{i=1}^{m}x_i\log x_i - \sum_{i=1}^{m}y_i\log y_i}{[(N_x + N_y)\log(N_x + N_y)] - N_x\log N_x - N_y\log N_y}$;

мера Барони-Урбани $M_{xy} = \dfrac{\sum_{i=1}^{m}\min(x_i, y_i) + \sqrt{[\sum_{i=1}^{m}\min(x_i, y_i)][\sum_{i=1}^{m}\max(s_i) - \max(x_i, y_i)]}}{\sum_{i=1}^{m}\min(x_i, y_i) + \sqrt{[\sum_{i=1}^{m}\min(x_i, y_i)][\sum_{i=1}^{m}\max(s_i) - \max(x_i, y_i)]}}$,

где N_x и N_y – суммы элементов векторов **x** и **y**, $\max(s_i)$ – максимальное значение i-го

[8] Другие его названия: индекс Ренконена, процентное подобие, коэффициент общности, индекс Штейнгауза, количественная мера сходства Чекановского и т.д.

показателя на всех обследованных объектах.

Если значения x_i и y_i принимают значения (0, 1), то мы имеем многочисленный класс мер расстояний, измеряемых в бинарном пространстве хеммингового типа. Таков, в частности, набор коэквивалентных индексов, которые по существу отражают долю общих биологических видов в среднем видовом богатстве сравниваемых выборок:

- Жаккара $K_{xy} = a/(a + b + c)$;
- Съеренсена $K_{xy} = 2a/(2a + b + c)$;
- Охаи (Ochiai) $K_{xy} = a/[(a + b)(a + c)]^{0.5}$,

где a – число совпадающих элементов, $b + c$ – число элементов, уникальных для **x** и **y**.

В ряде работ (Миркин, Розенберг, 1978) делаются попытки оценить, какие меры "завышают" или "занижают" сходство между сообществами и каким коэффициентам следует отдать предпочтение в работе. Однако вряд ли имеет смысл акцентировать эту сторону проблемы, так как единственным критерием адекватности оценок является устойчивость последовательностей агрегирования объектов на основании меры сходства в более крупные таксоны, иерархические деревья и проч.

Примеры многомерных пространств иного типа, применяемые в экологии и географии, подробно анализируются Ю.Г. Пузаченко (2004). Нам же важно подчеркнуть, что мера сходства/расстояния является типичным искусственно сконструированным собирательным понятием, отражающим латентные (т. е. скрытые, принципиально не измеряемые инструментально) свойства реальных объектов и включающим некоторые, выработанные практикой разумные критерии для качественного сопоставления этих объектов (P. Legendre, L. Legendre, 1998).

Поскольку любая из перечисленных индексов метрического или неметрического сходства является такой же мерой близости векторов **x** и **y**, что и коэффициент корреляции, для оценки их статистической значимости можно использовать рандомизационную процедуру, описанную в разделе 3.1. Однако, поскольку "дьявол таится в деталях", рассмотрим нетривиальные особенности этой задачи на примерах.

Предположим, нам необходимо выяснить, сформирован ли видовой состав донных сообществ верхнего (**x**) и нижнего (**y**) участков р. Сок [пример П2] из одного и того же пула видов. Соотношение оценок видового богатства макрозообентоса можно представить своеобразной таблицей сопряженности:

Число видов, обнаруженных во всех пробах		Нижний участок (44 пробы)		Итого:
		Найдено	Отсутствует	
Верхний участок (51 проба)	Найдено	a = 88	c = 102	190
	Отсутствует	b = 86	d = ?	86
Итого:		174	102	276

Частоты встречаемости каждого из 276 видов в пробах, взятых на каждом участке, составили пары значений x_i и y_i, которые были помещены в файл и использованы для расчета индексов, использующих количественные данные. Как и в большинстве предыдущих примеров, расчеты выполнялись в двух направлениях:

- оценка бутстреп-распределения и нахождение доверительных интервалов изучаемой меры сходства двух сообществ (мы не считаем, что их знание принесет здесь самостоятельную пользу, но интервальное оценивание всегда предпочтительнее точечного);
- формирование статистического распределения индексов при справедливости нулевой гипотезы (т.е. что оба сообщества образованы путем случайного извлечения из некого регионального пула видов) и оценка вероятности того, что наблюдаемый уровень сходства видового состава сообществ превысит случайный характер возможных совпадений.

Коды для расчета в статистической среде R представлены в дополнении к разделу. Значения мер сходства могли бы быть получены с использованием различных функций, представленных в пакетах vegan, labdsv или BiodiversityR, однако в методических целях

мы написали собственную функцию similary(). Напомним, что функция sample(X, replace = F) выполняет простую случайную перестановку значений вектора X и используется для рандомизации, тогда как sample(X, replace = T) осуществляет случайную выборку с возвратом и применяется для нахождения случайной комбинации порядковых номеров видов при бутстрепе. Результаты расчетов, выполненные на основе 1000 итераций рандомизации или бутстрепа, представлены в табл. 5.1.

Таблица 5.1. Результаты статистического анализа различных индексов биотического сходства участков р.Сок (ДИ – доверительный интервал, найденный методом процентилей)

Использованные индексы	Эмпирическая величина	Найденное бутстепом		При справедливости H_0		
		Смещение	95% ДИ	среднее	95% ДИ	*p*-значение
На основе количественных значений признаков						
Брея–Кёртиса	42.99	-0.31	33.88 ÷ 51.22	30.37	26.24 ÷ 35.58	0.001
Мориситы-Хорна	56.42	-0.19	47.31 ÷ 64.47	45.52	40.23 ÷ 51.48	0.001
Барони-Урбани	27.38	-0.02	20.42 ÷ 35.22	18.01	15.10 ÷ 21.76	0.001
На основе данных наличия/отсутствия						
Жаккара	31.88	-0.24	26.44 ÷ 37.31	49.11 30.11 *)	45.01 ÷ 53.58 25.95 ÷ 34.81	0.001 0.249
Съеренсена	48.35	-0.14	41.83 ÷ 54.35	65.86 46.48 *)	62.07 ÷ 69.78 40.65 ÷ 51.64	0.001 0.275
Охаи	48.39	-0.11	41.84 ÷ 54.74	65.90 46.51 *)	62.14 ÷ 69.84 41.78 ÷ 51.69	0.001 0.263

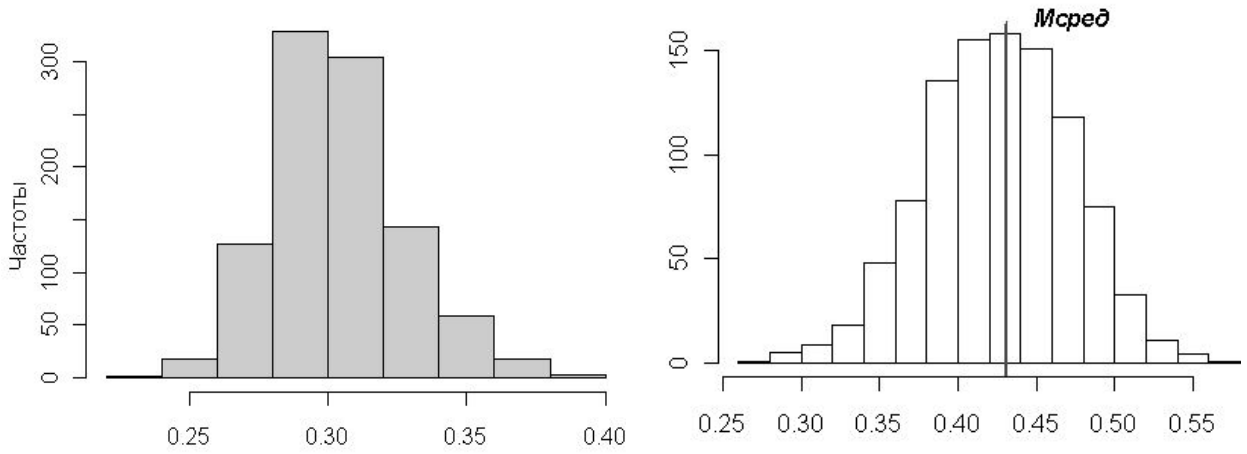
*) - во второй строке приведены результаты рандомизационного теста после добавления к эмпирической матрице блока "скрытых" видов с нулевой встречаемостью (d = 115).

При использовании мер сходства для количественных данных нулевая гипотеза о произвольном характере композиций видов отклоняется, поскольку корреляция популяционных плотностей в исходной выборке существенно выше, чем при случайном назначении. Однако важно не только подобрать экологическое объяснение, приличествующее результатам статистического анализа, но и ясно представлять себе математические тонкости проводимого теста.

Значения индексов этого класса почти полностью определяются различиями количественных показателей небольшой (5-10% от общего видового состава) группы ведущих таксонов-эврибионтов, которые слабо реагируют на изменение факторов окружающей среды, но доминируют как по численности особей, так и по частоте встречаемости. При случайных перестановках пары x_i - y_i таких таксонов разрушаются, а их представители входят в "мезальянс" с видами, имеющими малые популяционные характеристики. В качестве примера возьмем фрагмент обработанной таблицы из четырех видов. В крайнем правом столбце приведена одна из возможных перестановок средней численности особей нижнего участка по отношению к верхнему, но можно обратить внимание, насколько сильно при этом уменьшился индекс сходства:

Виды	Верх.Сок (эмпир.)	Нижн.Сок (эмпир.)	Нижн.Сок (рандом.)
Tanytarsus sp.	26	23	1
Cricotopus bicinctus	24	18	0
Ablabesmyia phatta	0	1	18
Acricotopus lucidus	1	0	23
Мера Брея–Кёртиса		44.1	2.1

Вероятность получить случайную комбинацию с еще более экстремальным значением меры Брея–Кёртиса в таких условиях невелика (см. рис. 5.1). Другой резон составляет фундаментальный вопрос социологии: «Является ли обоснованным судить о свойствах сообщества по группе социально аморфных, но многочисленных особей?»



а) При справедливости нулевой гипотезы

б) Бутстреп-распределение индекса сходства

Рис. 5.1. Распределение значений индекса Брея–Кёртиса сходства двух участков р. Сок, полученное рандомизацией (а) и бутстрепом (б); эмпирическое значение $M = 43$

Принципиально иная стратегия оценки видового подобия обнаруживается индексами, основанными на бинарных переменных (Шитиков и др., 2012). Они оценивают сходство по всему видовому составу, делая основной акцент на комплекс редких или трудно обнаруживаемых видов, встречающихся в единичных пробах с малой численностью (а таких видов часто оказывается существенное большинство). При сравнении двух участков р. Сок значения индексов сходства этого класса, полученные рандомизацией (т.е. при справедливости нулевой гипотезы), оказались существенно выше, чем на эмпирических данных – см. табл. 5.1 и рис. 5.2. Этот парадоксальный итог, что реальное сходство двух списков видов может оказаться меньше, чем при их случайном назначении, был самоуверенно объяснен нами (Шитиков, 2012) особенностями речного континуума вдоль продольного градиента, которые вызваны статистически значимым различием экологических условий реки в верхнем и нижнем течении.

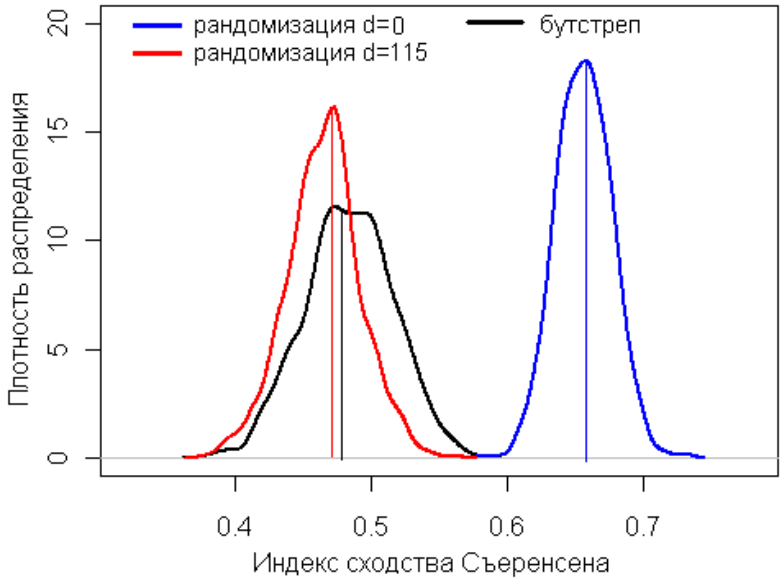


Рис. 5.2. Плотность бутстреп-распределения оцениваемого индекса Съеренсена и для нуль-моделей без добавления ($d = 0$) и с добавлением ($d = 115$) скрытых видов

При более зрелом размышлении стало очевидным, что причина этого парадокса – чисто математический эффект. Действительно, оказалось, что при отсутствии "нулевого хвоста" длиной d (т.е. видов, отсутствующих в обоих сравниваемых местообитаниях) вероятность создания новых пар (1, 1) в ходе перестановок будет несколько выше, чем их "разрушения". Тогда было решено учесть при рандомизации количество *невидимых* (или

скрытых за "линией занавеса" Ф. Престона) общих видов, которые потенциально могли бы встретиться при углублении мониторинговых исследований.

В большинстве случаев точная оценка *полного* видового богатства изучаемых сообществ требует дорогостоящих и масштабных исследований, что неизбежно ввиду биологической и пространственно-временной неоднородности природной среды. Привлекательной альтернативой ответа на вопрос "как много видов может встретиться?" является непараметрический статистический анализ с использованием ограниченных повторностей наблюдений (обзор возможных используемых методов – см. Шитиков и др., 2009). В частности, основная идея экстраполяции числа видов методом "складного ножа" (jackknife) заключается в расчете потерь числа видов при удалении одной из произвольных проб.

Пусть m – количество независимых выборок, взятых в изучаемой природной среде, а $\hat{S} = S_{obs}$ – число видов, обнаруженное во всех этих пробах. Если случайным образом исключить одну из выборок, то оставшееся видовое богатство на основе m - 1 измерений будет равно $\hat{S}_{-i} = S_{obs} - q_{1i}$, где q_{1i} – число уникальных видов, встретившихся только в i-й пробе. Тогда путем перебора всех значений $i = 1, \ldots, m$ можно найти значение суммарных потерь Q_{1*}, после чего рассчитать статистическую оценку видового богатства

$$\hat{S}^1_m = S_{obs} + Q_{1*}(m-1)/m$$

и её дисперсию

$$\mathrm{var}(\hat{S}^1_m) = \frac{1}{m}\sum_{i=1}^{m}(q_{1i} - Q_{1*}/m)^2 .$$

Эта оценка известна как "складной нож первого уровня" (Jackknife-1) и используется для компенсации статистического смещения оцениваемого параметра S_{obs} порядка $1/m$.

Для оценки числа скрытых видов удобно воспользоваться функцией specpool() из пакета vegan. Исходными данными для ее использования является упорядоченная по хронологии или иному ключу таблица числа особей, где в строках представлены пробы, а в столбцах – обнаруженные виды. Полное экстраполированное число видов макрозообентоса р. Сок составляет (276 + 115) = 391[9].

Этот расчет дает нам основания увеличить размерность векторов **x** и **y** с 276 до 391 путем добавления нулевых значений и повторить рандомизационный тест. Индексы сходства для бинарных признаков оказались в диапазоне доверительных интервалов нулевой модели и, следовательно, мы можем с уверенностью сказать, что видовой состав макрозообентоса в верховьях и в низовьях один и тот же, т.е. река представляет единое сообщество. Новый вариант анализа свидетельствует уже о *нейтральности* донного сообщества, в котором представленность видов подчинено случайным флуктуациям в однородных условиях среды, а взаимодействия между видами отсутствуют.

Мы не без оснований остановились на подробном разборе этого несложного примера, который ярко показывает, как часто статистический вывод может легко перейти в свою противоположность, если изменить некоторые исходные предпосылки, формулу тестового критерия или даже технику расчетов. Рандомизация – не исключение из этого общего правила, поэтому исследователю необходимо предварительно тщательно оценить состав ограничений, с учетом которых будут перемешиваться данные, и задать механизм перестановок, адекватный поставленной задаче.

К разделу 5.1:

```
source("similary.r")    # Загрузка из файла функции расчета метрик сходства двух векторов x-y
source("print_rezult.r")  # Загрузка функций вывода результатов
#     ---- Функция выполнения бутстреп-процедуры заданное bootN число раз
```

[9] Заинтересованный читатель может обратить внимание, что в комментариях к функции specpool() представлен пример оценки достоверности экстраполяции с использованием кросс-проверки

```
simci <- function(data , bootN, method)
    { boots <- numeric(bootN)   #  Вектор для накопления бутстрепированных значений
      vyb <- replicate(bootN, sample.int(ncol(data), replace=TRUE))
      boots <- sapply(1: bootN, function (i) {l<-vyb[,i]; similary(data[,l],method=method)})
  return (BootRes(boots, similary(data,method=method))) }  # Вывод результатов
# Выполнение бутстрепа
TT <- read.table("Dis.txt",header=T,sep="\t")  #  Загрузка численностей видов для x и y
TS <- t(TT) # Транспонируем матрицу видов (виды по столбцам, участки по строкам)
simci(TS,1000,4) ; simci(TS,1000,5) ; simci(TS,1000,6)   #   Для количественных переменных
simci(TS,1000,4) ; simci(TS,1000,5) ; simci(TS,1000,6)   #   Для бинарных переменных
#   ----------  Функция выполнения рандомизации заданное permutations число раз
simP <- function(data , permutations, method)  {
  empar <- similary(data,method=method) ; boots <- numeric(permutations)
  for (i in 1:permutations)
       #   Каждый раз заменяем вектор y случайной перестановкой из его значений
       boots[i] <- similary(rbind(data[1,], sample(data[2,], replace=FALSE)),method=method)
   return (RandRes (empar, boots, permutations)) }  #    Вывод результатов
#   Выполнение рандомизации
simP(TS,1000,4) ; simP(TS,1000,5) ; simP(TS,1000,6)   #   Для количественных переменных
#   Оценка числа видов, скрытых за занавесом, методом складного ножа
#   Подготавливаем и загружаем из файла матрицу с численностями особей T,
#   где в строках – гидробиологические пробы, а в столбцах – обнаруженные виды
      library (vegan)
      specpool(T, index = "jack1")  # Используем только метод складного ножа 1
#  Результат:  d = 115
#   Добавляем справа к матрице наблюдений  115 столбцов с нулевой численностью
TSD <- cbind(TS, matrix(replicate(230,0), nrow=2))
simP(TSD,1000,1) ; simP(TSD,1000,2) ; simP(TSD,1000,3)   #   Для бинарных переменных
```

5.2. Непараметрический дисперсионный анализ матриц дистанции

Результаты мониторинга природного объекта, представленные в виде m выборок, обычно можно разделить на r групп, например, в соответствии со схемой зонирования водотока, где выполнялись гидробиологические пробы. Тогда, в соответствии с моделью дисперсионного анализа, общую изменчивость популяционной плотности N_i каждого i-го вида можно разложить на компоненты: $\mathrm{Var}\ N_i = \mathrm{Var}\ \tau + \mathrm{Var}\ \varepsilon$, где $\mathrm{Var}\ \tau$ – вариация, обусловленная влиянием группирующего фактора, $\mathrm{Var}\ \varepsilon$ – изменчивость, связанная с воздействием неконтролируемых (в том числе, случайных факторов).

Однако, если количество зарегистрированных видов велико (300-400 в гидробиологических примерах), то дисперсионный анализ не дает возможности оценить совокупную изменчивость всего таксономического комплекса изучаемого сообщества под воздействием группирующего фактора. Один из приемов избежать "проклятия размерности" – выполнить дисперсионный анализ симметричной матрицы **D** размерностью $m \times m$, элементами которой d_{ij} являются коэффициенты расстояния между каждой парой выполненных проб (см. рис. 5.3).

М. Андерсон (Anderson, 2001) предложил метод, известный как непараметрический дисперсионный анализ (npMANOVA), осуществляющий разложение многомерной изменчивости, заключенной в матрице расстояний, в соответствии с уровнями влияния изучаемых факторов. В первом приближении, как это показано на рис. 5.3, можно рассчитать общую SS_T и внутригрупповую SS_W суммы квадратов d_{ij}, после чего оценить их дисперсионное отношение:

$$SS_T = \frac{1}{m}\sum_{i=1}^{m-1}\sum_{j=i+1}^{m} d_{ij}\ ;\ \ SS_W = \frac{1}{r}\sum_{i=1}^{m-1}\sum_{j=i+1}^{m} d_{ij}\omega_{ij}\ ;\ F = \frac{SS_W(m-r)}{SS_T(r-1)}, \qquad (5.1)$$

где ω_{ij} равны 1, если выборки i и j принадлежат одной группе, и 0 в противном случае.

Рис. 5.3. Схема группировки матрицы расстояний

Однако любая метрика расстояния, принятая в качестве модели отношений между точками данных, не вполне вписывается в общую концепцию классического дисперсионного анализа, оценивающего вариацию случайных измерений относительно групповых средних. Поэтому был предложен (McArdle, Anderson, 2001) другой способ вычисления тестовой статистики с использованием концепций линейной модели ANOVA.

При применении этого подхода на основе исходной матрицы расстояний **D** с использованием преобразований Гувера рассчитывается матрица **G**, описывающая "центроиды", т. е. центры распределения значений d_{ij} в пределах каждой группы. Другая проекционная матрица **H** (или "шляпа" hat) вычисляется с использованием вектора ортогональных контрастов, связанных с уровнями группировочных факторов. Тогда тестовая статистика для проверки нулевой гипотезы об отсутствии различий между r группами или псевдо F-критерий находится как

$$F = \frac{tr(\mathbf{HGH})(m-r)}{tr[(\mathbf{I}-\mathbf{H})\mathbf{G}(\mathbf{I}-\mathbf{H})(r-1)}, \qquad (5.2)$$

где tr – сумма диагональных элементов матрицы, **I** – единичная матрица. В несколько преобразованном виде выражение для псевдо-F-критерия может быть использовано в многофакторном дисперсионном анализе матриц дистанций.

Можно отметить, что при особой конфигурации данных и использовании евклидова расстояния, обе формулы (5.1) и (5.2) становятся эквивалентными и воспроизводят традиционное F-отношение Фишера. Но, в общем случае, распределение многомерной псевдо-F-статистики при справедливости нулевой гипотезы теоретически неизвестно, поэтому для его восстановления используется рандомизационная процедура. Ее характер уже вполне знаком читателю по примерам, подробно описанным выше: исходные выборки многократно случайным образом перемешиваются относительно выделенных групп. Практическая стратегия здесь состоит в том, чтобы получить большое подмножество значений F^*, извлеченных хаотично, независимо и с равной вероятностью из нуль-распределения $\boldsymbol{F^*}$ для всех возможных перестановок. P-значение, соответствующее верной нулевой гипотезе, вычисляется как доля значений F^*, которые больше или равны значению F, найденному на реальных данных.

Здесь следует добавить, что npMANOVA дает возможность использовать три алгоритма перестановки при реализации рандомизационного теста: 1) перемешивание исходной таблицы данных, 2) перестановка остатков для сокращенной модели и 3) перестановка остатков для полной модели. Хотя эти три подхода дадут очень похожие результаты, в статье (Anderson, ter Braak, 2003) представлено резюме по их применению.

Используем пример [П2], рассмотренный в разделе 4.4 и основанный на двухступенчатом гнездовом разделении 120 гидробиологических проб макрозообентоса р. Сок. Было выделено три географически обособленных участка водотока, внутри которых данные пространственно распределялись по 4 станциям в разных створах реки.

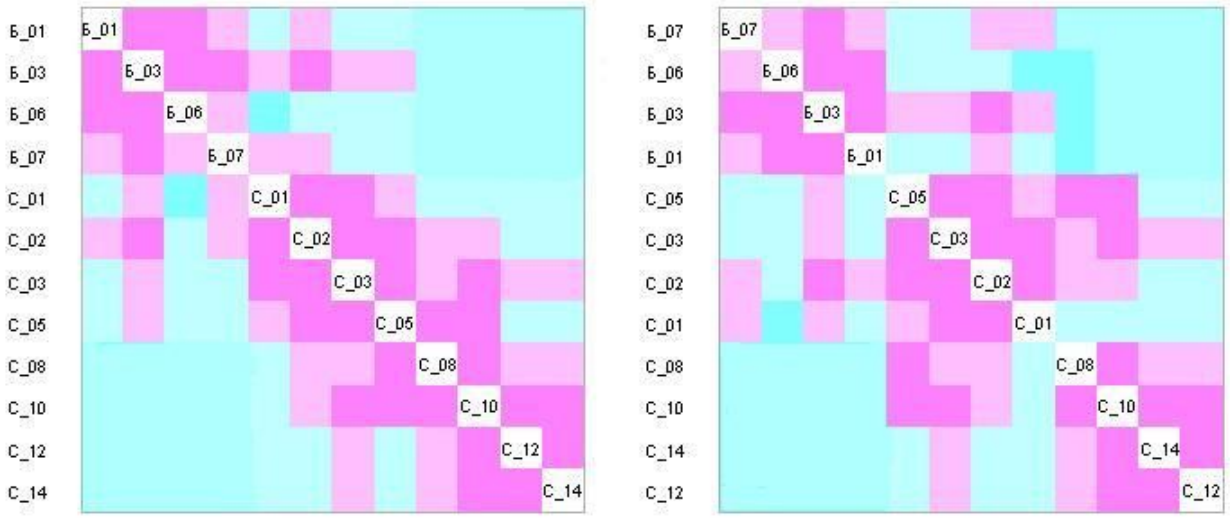
Сформируем две матрицы дистанций **D** размерностью 120 × 120, оценивающие различие таксономической структуры проб в пространстве 363 видов, численность обнаруженных экземпляров которых предварительно логарифмировалось. В качестве конкретных дефиниций метрик будем использовать количественную меру Брея–Кёртиса и индекс Жаккара (во втором случае таблица наблюдений интерпретируется в режиме наличия/присутствия вида: binary=TRUE).

Дисперсионный анализ npMANOVA проведем в среде R с использованием функции adonis(…) из пакета vegan – см. табл. 5.2. Формула линейной модели и включение параметра стратификации (strata) обеспечили условия выполнения иерархического гнездового (nested) ANOVA, т.е. изменчивость видовой структуры на станциях оценивалась только внутри каждого участка.

Таблица 5.2. Результаты дисперсионого анализа матрицы дистанций методом непараметрического MANOVA с использованием функции adonis()

Факторы	Степени свободы *df*	Сумма квадратов	Средние квадраты	*F*-критерий	R^2	P = Pr(>F)
Компоненты матрицы дистанций - мера Брея–Кёртиса						
A (Участок)	1	3.14	3.14	9.37	0.073	**0.002**
B{**A**} Станция	3	1.72	0.573	1.57	0.036	**0.002**
Остаток (error)	115	41.8	0.363		0.891	
Компоненты матрицы дистанций – индекс Жаккара						
A (Участок)	1	2.62	2.62	6.44	0.051	**0.001**
B{**A**} Станция	3	1.78	0.595	1.46	0.035	**0.001**
Остаток (error)	115	46.7	0.406		0.913	

Результаты анализа, представленные в табл. 5.2, свидетельствуют о высокой статистической значимости отличий видовой структуры донных сообществ на выделенных участках. Визуально это легко видно на диаграмме (рис. 5.4), где представлена матрица расстояний, тональность раскраски ячеек которой соответствует уровню сходства. Впрочем, статистически значима (хотя в меньшей мере) и вложенная вариация таксономического состава между станциями внутри каждого участка.



а) Матрица дистанций б) Отсортированная матрица дистанций

Рис. 5.4. Диаграмма сходства станций р.Сок в осях двух главных координат

Функция adonis (…) на основе типа I суммы квадратов и для каждого терма модели рассчитывает также статистику, условно названную R^2, которая оценивает силу влияния факторов. Ориентируясь на значения статистик F и R^2, можно сделать вывод, что

использование количественных индексов предпочтительнее для сравнения видовых структур, чем индексов, основанных на признаках наличия/присутствия.

Как и многие другие версии дисперсионного анализа, тест npMANOVA, оценивающий различие в мерах положения отдельных групп наблюдений в многомерном пространстве, также чувствителен к неоднородности дисперсии в группах. Если сравниваемые сообщества сильно отличаются между собой по β-разнообразию, то причинами отклонения нулевой гипотезы могут явиться как сдвиги центров групп между собой, так и различия в уровне внутригрупповой изменчивости (или оба фактора вместе).

Другая процедура М.Андерсона betadisper(…) выполняет анализ внутригрупповой однородности мер расстояний матрицы **D** и является многомерным аналогом теста Левене (Levene – одномерную версию теста см. в разделе 2.6) на однородность дисперсий в группах. Последняя версия этого алгоритма (Anderson, 2006) сводится к следующему:

- вычисляются оси главных координат PCoA1, PCoA2, PCoA3 и т.д. матрицы **D** (как это сделать, описывается в главе 6);
- находятся "центроиды" (или пространственные медианы) выделенных групп в этих коодинатах – см. рис. 5.5;
- внутригрупповые дисперсии определяются как сумма евклидовых расстояний между центроидом и точками, соответствующими выполненным наблюдениям, для каждой анализируемой группы.

Чтобы проверить, отличаются ли средние расстояния до центроидов в одной или более групп от остальных, выполняется обычный анализ дисперсионных отношений ANOVA, причем статистическая значимость *F*-критерия группы может быть рассчитана как параметрическими методами, так и с использованием рандомизационного теста. Во втором случае процедура permutest.betadisper(…) заданное число раз перемешивает остатки линейной модели, чтобы сгенерировать распределение перестановочного *F*-отношения при справедливости нулевой гипотезы об отсутствии между группами отличий в оценках дисперсии.

В нашем примере гипотеза об однородности вариаций внутригруппового сходства по участкам в целом отклоняется при $F = 6.89$ ($p = 0.002$). Выполним однако множественные парные сравнения оценок величины дисперсии и найдем статистическую значимость отличий β-разнообразия на отдельных участках:

- Байтуган – Сок (верхний участок): $p = 0.0191$ (пермутационный тест $p = 0.019$);
- Байтуган – Сок (нижний участок): $p = 0.001$ ($p = 0.002$);
- Сок (верхний участок) – Сок (нижний участок): $p = 0.093$ ($p = 0.097$).

На основе этих результатов легко также сделать вывод, что макрозообентос р. Байтуган представляет более компактное и внутренне однородное сообщество – см. рис. 5.5. В противоположность этому, на двух участках основного водотока р. Сок β-разнообразие отличается между собой, поскольку внешние воздействия антропогенного характера и разнообразие природных условий определяют более сильную флуктуацию видовой структуры.

Результаты дисперсионного анализа npMANOVA (табл. 5.2) привели нас к выводу, что гидробиологические пробы взяты из разных генеральных совокупностей. Однако было бы интересно выяснить, какие именно группы биотопов в наибольшей мере определяют эти различия и "выбиваются из общей колеи". Следует напомнить, что множественные сравнения необходимо проводить специальными тестами, иначе нам не избежать ошибки I рода нахождения различий там, где их нет (т.е. нулевая гипотеза неверно отвергнута статистическим критерием).

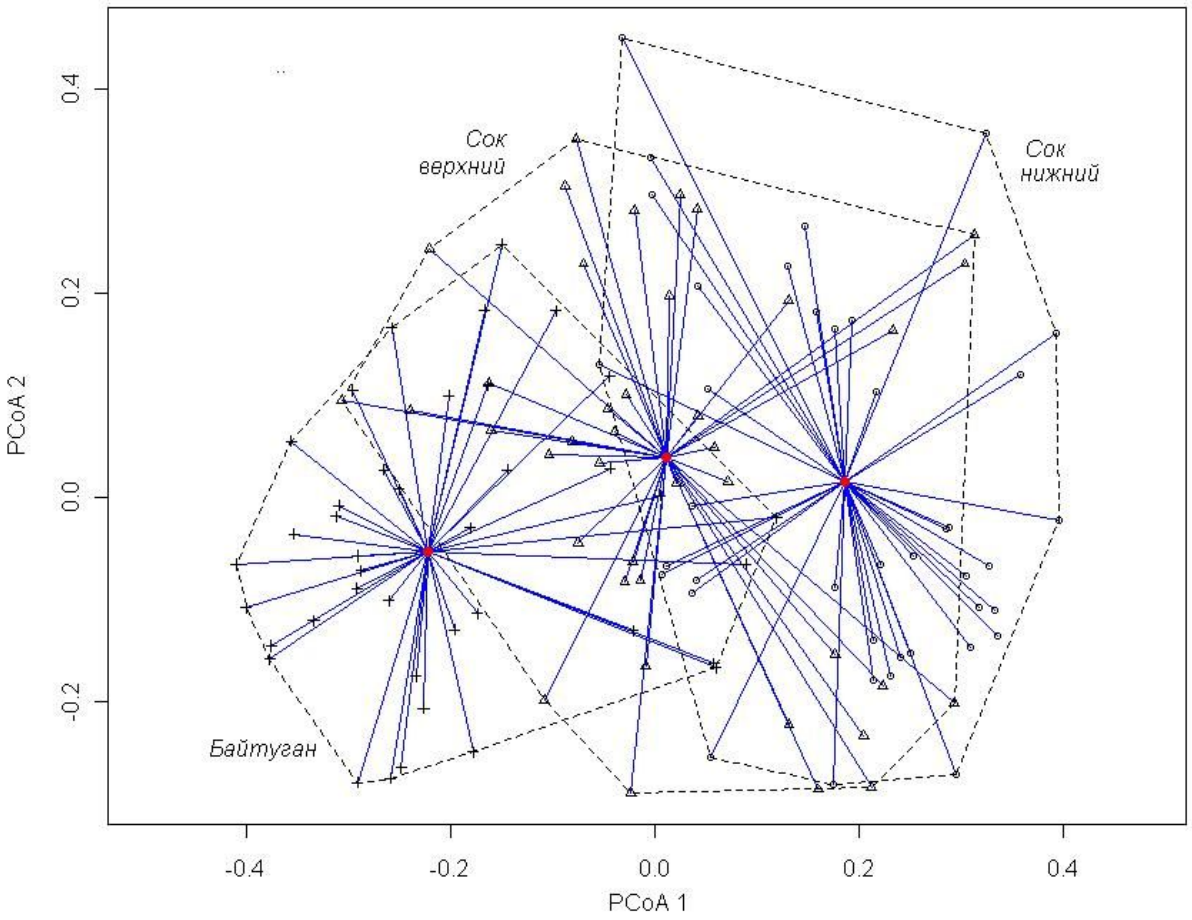


Рис. 5.5. Ординационная диаграмма сходства участков р. Сок в осях двух главных координат

Наиболее распространённым и рекомендуемым в литературе является тест Тьюки, использующий критерий "подлинной" значимости (Honestly Significant Difference, HSD). HSD задает наименьшую величину разности математических ожиданий в группах, которую можно считать значимой, и ее одновременные доверительные интервалы – см. рис. 5.6. Метод, в отличие от схемы Бонферрони (см. раздел 2.5), не конструирует одну комплексную гипотезу, а выполняет только попарные сравнения: если доверительные интервалы разности для анализируемой пары не включает 0, то частная H_0 отклоняется.

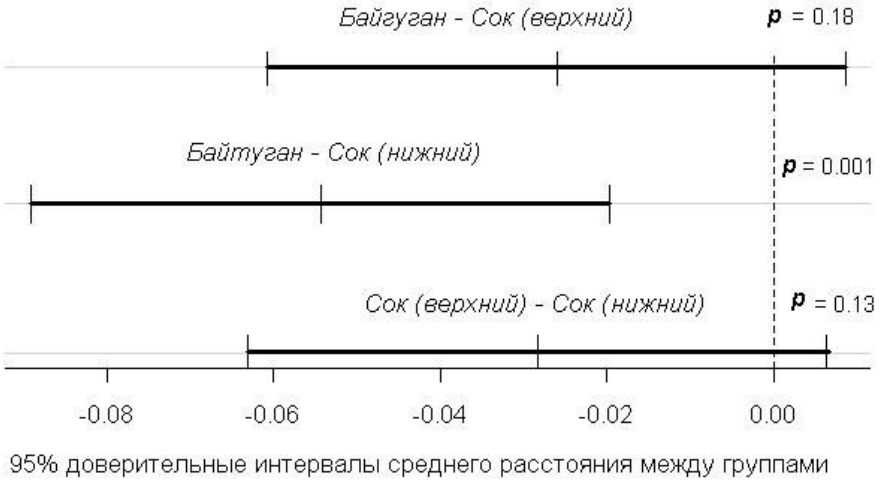


Рис. 5.6. Множественные сравнения участков р.Сок на основе видового сходства донных сообществ с использованием критерия HSD Тьюки

Тест Тьюки HSD реализуется в R для модели betadisper с использованием специальной функции TukeyHSD(…). Результаты, представленные на рис. 5.6, показывают, что различия, обозначенные npMANOVA, целиком относятся на счет территориально разобщенной пары "Байтуган - Сок (нижний участок) ", тогда как для остальных пар нулевая гипотеза не отвергается. Наконец, необходимо отметить, что критерий Тьюки (как и большинство других критериев парных сравнений) применяется в предположении, что дисперсии всех сравниваемых групп равны. Если вспомнить, что это условие не соблюдается, то внешнюю непоколебимость вывода о специфичности видового состава на разных участках реки уже можно подвергнуть сомнению.

К разделу 5.2:

```
# Загрузка данных из предварительно подготовленного двоичного файла (см. раздел 4.4.)
load(file="Sok.RData")
# attach делает «видимыми» имена столбцов таблицы
attach(TTB.Site) ; library(vegan)
River = as.factor(River)    # Главный фактор
num_Site = as.factor(num_Site)     # Вложенный фактор
#  Выполняем npMANOVA на основе индекса Жакара
adonis(formula = TTB ~ River + num_Site %in% River, data = TTB.Site,
                      strata= River, method="jac", binary=TRUE,  permutations = 999)
D <- vegdist(TTB)  #  Предварительно рассчитываем матрицу расстояний Брея-Кёртиса
#   А выполнить функцию Adonis()  можно и так:
adonis(formula = D ~ River + Site %in% River, data = TTB.Site, strata= River,
                          permutations = 999)
#   Создание модели – объекта betadisper
mod <- betadisper(D, River, type = "centroid")
#  Вывод таблицы дисперсионного анализа, и графиков в осях 1-2 и 1-3 главных координат
anova(mod) ; plot(mod);   plot(mod, axes = c(3,1)
#  Запуск перестановочного теста, вывод диаграммы «бокс с усами» для дисперсий
permutest(mod, pairwise = TRUE) ; boxplot(mod)
#   Выполнение теста Тьюки HSD, вывод графика с доверительными интервалами
(mod.HSD <- TukeyHSD(mod)) ;  plot(mod.HSD)
#   Построение диаграммы – «расцвеченной» матрицы расстояний
#   Суммирование численностей видов по повторностям для каждой станции и расчет средних
TTB.agr <- aggregate(TTB,list(TTB.Site$Site), FUN = mean) ; TTB.agrs <- TTB.agrs[,-1]
rownames(TTB.agrs)<- c("Б_01","Б_03","Б_06","Б_07","С_01","С_02","С_03","С_05",
                                                "С_08","С_10","С_12","С_14")
#   Формирование матрицы расстояний Брея-Кёртиса между 12 станциями
library(gclus) ; Da <- vegdist(TTB.agrs)
#   Подключение модуля с функцией  coldiss()  (Borcard et al., 2011) и вывод графика
source("coldiss.r") ; coldiss(Da,byrank=TRUE, diag=TRUE)
```

5.3. Тест Мантеля для оценки связи между многомерными структурами

Рассмотрим теперь проблемы оценки силы связи между двумя многомерными структурами данных, представленных матрицами расстояний. Эта задача возникает, например, когда необходимо проверить обоснованность следующих содержательных гипотез:

- связаны ли статистически значимой зависимостью географическое расстояние между местообитаниями и экологическое сходство их видовых составов;
- имеется ли взаимосвязь между двумя группами организмов, обитающих в одних и тех же биотопах (например, включающих растения и организмы беспозвоночных);
- имеются ли изменения в видовой структуре сообществ до и после внесения возмущения.

Метод Мантеля или "соотношение квадратов" (quadratic assignment – Mantel, 1967) проверяет гипотезу H_0, что расстояния (или близости) между объектами в матрице **A** независимы от расстояний (или близостей) между теми же самыми объектами в матрице

B. Тест является альтернативой регрессии при анализе зависимости первой матрицы расстояний от второй, вычисляя, по сути, коэффициент корреляции между ними, не нуждаясь при этом в каких-либо строгих статистических предположениях о характере распределения данных. В связи с этим, метод позволяет найти линейные отношения между матрицами, составленными на основе меры расстояния любой природы.

Предположим, что обе матрицы расстояний **A** и **B** относятся к одному и тому же фиксированному набору объектов. Если найти сумму элементов в матрице **Z = AB**, которая является произведением обоих сравниваемых матриц расстояний **A** и **B**, при этом исключая элементы на главной диагонали, то получим *Z*-статистику Мантеля, т. е. $Z = \sum_{i=1}^{n-1}\sum_{j=i+1}^{n} a_{ij}b_{ij}$. Стандартизованная статистика Мантеля *r* изменяется от +1 до –1 и соответствует коэффициенту корреляции Пирсона *r* между матрицами расстояний **A** и **B**.

Для оценки нулевой гипотезы H_0 об отсутствии зависимости между матрицами **A** и **B** может быть использована *аппроксимация Мантеля* (Mantel, 1967), которая трансформирует *Z*-статистику в *t*-значение, распределение которого асимптотически приближается к нормальному. Метод асимптотической аппроксимации Мантеля дает хорошие результаты для больших наборов данных.

Если элементы матриц **A** и **B** рандомизированы (т. е. представлены случайными наборами чисел), то процессы, порождающие сравниваемые матрицы, являются независимыми. В соответствии с этим, оценка нулевой гипотезы H_0, не связанная с какими-либо априорными предположениями, может быть получена с помощью перестановочного теста: в обеих матрицах расстояний случайным образом меняется порядок рядов и строк. Если процедуру перестановки повторять многократно (например, B = 1000 раз), то можно смоделировать распределение значений статистики **Z***. Значимость связи между эмпирическими матрицами сходства оценивается как выход статистики Z за пределы правостороннего доверительного интервала распределения **Z***, а вероятность p ошибки I рода находится по формуле $p = (1 + b)/(1 + B)$, где b – число итераций, когда модельное значение Z_{sim} оказывается больше эмпирического Z_{obs}.

В разделе 4.5 рассматривалась регрессионная модель зависимости надземной биомассы растений на 159 участках в дельте реки Волги от химического состава почвы (пример [П3]). Схема расположения точек наблюдений и распределение обилия травянистого покрова представлена на рис. 5.7.

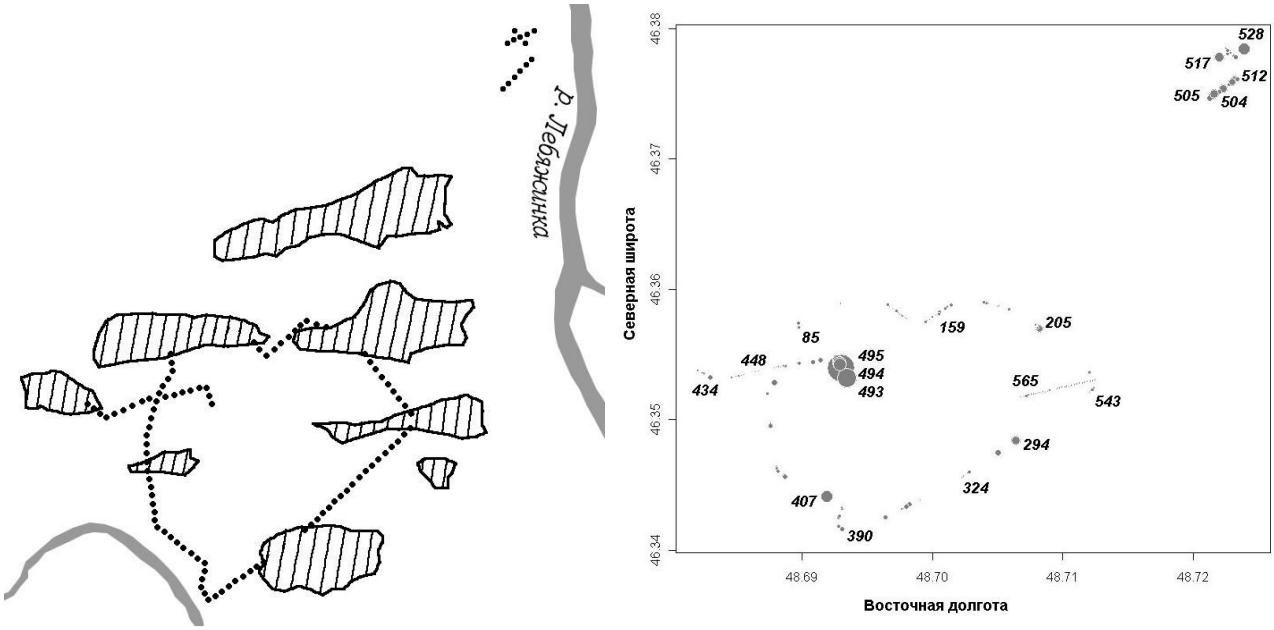


Рис. 5.7. Схемы расположения площадок для взятия геоботанических проб в дельте р. Волга (слева показаны номера площадок, а размер кругов соответствует общей биомассе растений)

На основе этих геоботанических исследований подготовим три таблицы данных:

- состав травянистого покрова и биомасса (г/м^2) отдельных видов растений (тростника, рагозы, пырея, череды, алтея и др.) – таблица **A** размерностью 23×159;
- катионный и анионный состав почвы и другие параметры среды – **B** (20×159);
- географические координаты точек отбора проб (северная широта и восточная долгота – **G** (2×159).

Для того, чтобы обеспечить улучшение статистических свойств выборок и разрешающую способность методов последующего анализа, исходные таблицы подвергают трансформации и стандартизации. Таблицу обилия видов **A** преобразуем во формуле Хеллингера (Legendre, Gallagher, 2001): $a'_{ij} = \sqrt{a_{ij} / \sum_j a_{ij}}$, т.е. обилие каждого *i*-го вида делится на общую сумму его обилия на всех площадках и из этой доли извлекается квадратный корень. Поскольку все виды оказываются представленными в единой шкале (0 ÷ 1), снижается удельное влияние таксонов с высокой популяционной плотностью и повышается внимание к комплексу редких видов.

Таблицу факторов среды **B** подвергнем стандартизации $b'_{ij} = (b_{ij} - \bar{b}_i) / s_i$, т.е. отклонение каждого показателя от его среднего значения разделим на стандартное отклонение. Такое *z-преобразование* переменных в единую шкалу обеспечивает соизмеримость выборок с различными средними и дисперсиями в рамках их совместной обработки, и при этом не происходит какой-либо потери информации. На основе преобразованных таблиц **A'** и **B'** вычислим матрицы евклидовых расстояний $\mathbf{D_A}$ и $\mathbf{D_B}$ размерностью 159×159 каждая.

Таблицу географических координат **G** сначала преобразуем в числовую форму, а затем по формуле *Haversine* сформируем матрицу натуральных расстояний $\mathbf{D_G}$ (км) размерностью 159×159 между каждой парой точек наблюдений.

Результаты оценки достоверности статистической связи между комплексами переменных **A, B** и **G** (табл. 5.3) с использованием теста Мантеля показывают, что отсутствует линейная связь между высотой площадок и показателями ионного состава почвы, заданных матрицей **B**, и их географическими координатами **G**. Следовательно план геоботанических исследований был в достаточной степени рандомизирован относительно поставленной задачи и влияние пространственного градиента на изменчивость видовой структуры фитоценозов сказываться не будет. Отметим, что доверительные интервалы корреляции Мантеля и соответствующие ему *p*-значения в табл. 5.3 были получены на основе перестановочного теста после 999 итераций.

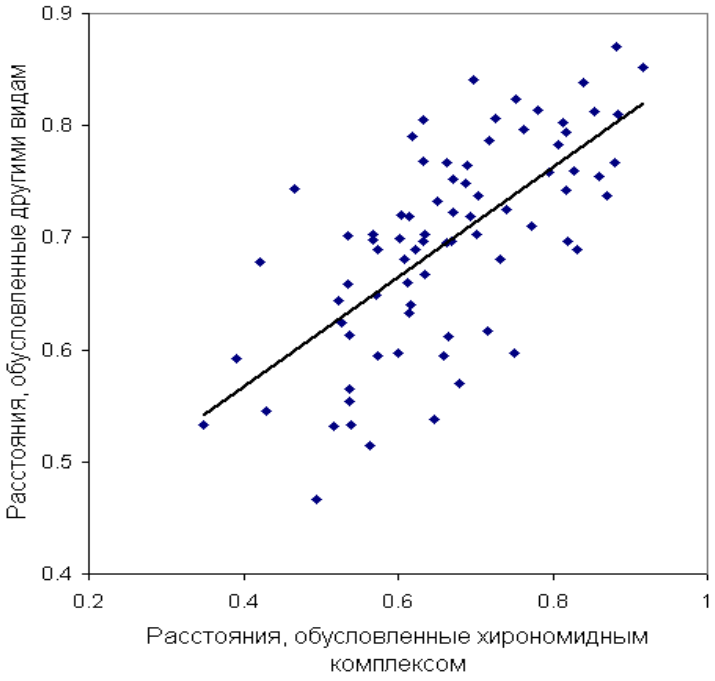
Таблица 5.3. Результаты теста Мантеля для оценки статистической значимости связи обилия видов травянистого покрова с факторами среды и пространственным расположением

Сравниваемые матрицы дистанций	Статистика Мантеля *r*	95% доверительные интервалы *r*	*p* = Pr(*r*>*0*) при рандомизации
{Обилие видов **A**} ~ {Факторы среды **B**}	0.248	0.198 ÷ 0.298	**0.001**
{Обилие видов **A**} ~ {Расположение в пространстве **G** }	0.182	0.146 ÷ 0.218	**0.001**
{Факторы среды **B**} ~ {Расположение в пространстве **G** }	0.026	- 0.041 ÷ 0.093	0.197
Связь между тремя матрицами **A, B** и **G** одновременно	0.247	0.200 ÷ 0.294	**0.001**

С использованием теста Мантеля может быть выполнен не только анализ пространственной изменчивости, но и оценка связей между компонентами экосистемы. Например, исследователь может задаться вопросом, изменятся ли результаты ординации местообитаний, если использовать некоторое подмножество списка видов: в частности,

ограничиться только одной систематической группой или взять данные за разные экспедиционные периоды. Поскольку в основе всех описываемых здесь методов лежит расчет матрицы сходства или корреляции, то в математической плоскости эта проблема сводится к анализу: статистически значимы ли различия между двумя произвольными симметричными матрицами расстояний **A** и **B**.

Оценим, имеется ли статистическая взаимосвязь между характером распределения 165 видов семейства хирономид (Chironomidae) и 210 остальных видов макрозообентоса по 13 створам рек Байтуган-Сок (см. пример [П2], рассмотренный также в разделах 4.4 и 5.2). Матрицы расстояний **A** и **B** между местообитаниями рассчитаем с использованием формулы Брея-Кёртиса. На рис. 5.8 представлен характер взаимосвязи между этими матрицами: каждая из точек биплота соответствует расстояниям между каждой парой местообитаний для двух композиций видов. Если эта зависимость имеет статистически значимый линейный характер, то все семейства макрозообентоса в одинаковой мере реагируют на биотопическую изменчивость по руслу водотока.



Эмпирические значения
Статистика Мантеля $Z_{obs} = 19{,}6$
Стандартизированная статистика Мантеля $r = 0{,}657$;

Аппроксимация Мантеля
- t-критерий = 5,84;
- $p = 0{,}001$

Перестановочный тест Монте-Карло
Параметры рандомизированного распределения:
- среднее $Z_{sim} = 19{,}16$:
- минимум $Z_{sim} = 18{,}95$;
- максимум $Z_{sim} = 19{,}43$;

Число перестановок $N = 999$;
Число значений с $Z_{sim} > Z_{obs}$ $n = 0$;
Число значений с $Z_{sim} < Z_{obs}$ – 999;
Вероятность p ошибки I рода = 0,001.

Рис. 5.8. Характер взаимосвязи между значениями матриц расстояний, обусловленных двумя подмножествами видов макрозообентоса (реки Байтуган-Сок, 13 станций наблюдений)

Представленные на рис. 5.8 результаты теста Мантеля показывают статистически значимую связь между характером распределения двух комплексов видов макрозообентоса по течению водотока. Сходные выводы могут быть получены при сравнении матрицы расстояний **A**, рассчитанной по полному списку видового состава макрозообентоса, с двумя другими матрицами:

- **B´** по хирономидному комплексу – $r = 0{,}756$, $t = 6{,}59$, $p \cong 0{,}0$;
- **B´´** по списку видов, не включающих хирономид – $r = 0{,}744$, $t = 6{,}50$, $p \cong 0{,}0$.

Эти результаты свидетельствует о тесной экологической связи между обоими компонентами макробентоценоза.

Тест Мантеля стал основой процедуры *множественной матричной регрессии* (Smouse et al., 1986; Manly, 2007), приводящей к модели следующего вида: $\mathbf{Y} = \beta_0 + \sum_{i=1}^{m} \beta_i \mathbf{A}_i + \mathbf{E}$, где **Y** – зависимая матрица; $\mathbf{A}_i$ – матрицы, независимые от **Y**; β_0 и β_i – свободный член и коэффициенты регрессионной модели; **E** – матрица ошибок

(остатков). Метод специально предназначен для анализа взаимосвязи матриц сходства или корреляции, у которых элементы, расположенные по главной диагонали, равны 1. Множественная процедура Мантеля реализует при этом обычную линейную регрессию, однако вследствие специфической структуры исходных данных для оценки параметров модели используются перестановочные методы Монте-Карло.

К сожалению, нам не удалось найти функцию R, реализующую множественную матричную регрессию, однако эти расчеты могут быть выполнены с использованием программ RTMant и RTMant1, разработанных Б.Манли в составе пакета RT (Randomization Testing). Другая и более удобная версия программы Multi_mantel разработана Л.Ревеллом (Revell) и свободно распространяется с ее исходными модулями на языке С на сайте http://anolis.oeb.harvard.edu/~liam/programs/.[10]

Оценим, какой вклад вносят отдельные таксономические группы макрозообентоса (на уровне семейств и подсемейств) в общую изменчивость донных сообществ по руслу рек Байтуган-Сок. Рассчитаем частные матрицы сходства, составленные из коэффициентов Брея–Кёртиса, для видов семейств Ephemeroptera ($\mathbf{A}_1$), Oligochaeta без *Nais* sp. ($\mathbf{A}_2$), Trichoptera ($\mathbf{A}_3$) и Coleoptera ($\mathbf{A}_4$), а также для отдельных таксонов хирономид Orthocladiinae ($\mathbf{B}_1$), Tanypodinae ($\mathbf{B}_2$), Chironomini ($\mathbf{B}_3$) и Tanytarsini ($\mathbf{B}_4$). В качестве зависимых матриц будем использовать аналогичные матрицы сходства, рассчитанные по всему комплексу видов (отдельно для хирономид и остальных таксонов). Результаты множественного матричного регрессионного анализа представлены в табл. 5.4.

Таблица 5.4. Параметры моделей множественной матричной регрессии для оценки вклада отдельных групп макрозообентоса в изменчивость донных сообществ речной экосистемы Байтуган–Сок

	Коэффициенты β	Стандартное отклонение	*t*-критерий Стьюдента	*p*-значение вероятности	Коэффициент детерминации (R^2) и критерий Фишера (F)
Модель **V** – виды, не относящиеся к семейству Chironomidae					
Свободный член	0,127				$R^2 = 0{,}552$ $F = 22{,}5$ ($p = 0{,}001$)
Ephemeroptera	***0,334***	0,041	8,13	***0,001***	
Oligochaeta	0,085	0,046	1,81	0,121	
Trichoptera	0,083	0,046	1,80	0,092	
Coleoptera	-0,007	0,031	-0,22	0,840	
Модель **C** – виды, относящиеся к семейству Chironomidae					
Свободный член	-0,040				$R^2 = 0{,}888$ $F = 144{,}8$ ($p = 0{,}001$)
Orthocladiinae	***0,491***	0,040	12,30	***0,001***	
Tanypodinae	0,014	0,024	0,58	0,670	
Chironomini	***0,381***	0,033	11,55	***0,001***	
Tanytarsini	***0,201***	0,035	5,64	***0,001***	

На основании результатов анализа можно сделать вывод, что при общей значимости моделей регрессии по критерию Фишера далеко не все таксономические группы вносят одинаковый вклад в объяснение вариации видовой структуры по руслу рек. В частности, для модели **V** статистически значимо отличается от 0 коэффициент регрессии для семейства Ephemeroptera, тогда как остальные группы макрозообентоса не оцениваются как значимые для объяснения вариации коэффициентов сходства участков. Для более однородного хирономидного комплекса (модель **C** в табл. 5.4) основные группы видов, за исключением Tanypodinae, вносят приблизительно одинаковый вклад в пространственную изменчивость видовой структуры.

[10] Л. Ревелл после наших контактов любезно разработал скрипт для расчета матричной регрессии в статистической среде R: http://phytools.blogspot.ru/2012/10/multiple-matrix-regression-with-mantel.html. Мы приводим текст этого скрипта в комплекте распространяемых файлов и заинтересованный читатель сможет самостоятельно выполнить расчеты, подобные табл. 5.4.

К разделу 5.3:

```
# Загрузка данных из предварительно подготовленного двоичного файла
load(file="Fito_Full.RData") ;  library(vegan)
# Формирование таблицы видов и преобразование ее в матрицу дистанций Хеллингера
Species <- Fito_Full[,3:25] ;  Species[Species<0] <- 0
SpecN <- decostand(Species, method="hellinger") ; D_spec <- dist(SpecN)
# Формирование таблицы факторов среды, преобразование ее в стандартизованный вид
Envir <- Fito_Full[,27:36] ; EnvN <- decostand(Envir, method="standardize")
#   Формирование матрицы евклидовых дистанций
D_env <- dist(EnvN)
#   Формирование таблицы географических расстояний, превращение координат в число
library(gmt) ; Geo <- Fito_Full[,37:38]
Geo$Yc <- deg2num(gsub(" ","", Geo$Yc , fixed=TRUE))
Geo$Xc <- deg2num(gsub(" ","", Geo$Xc , fixed=TRUE))
# geodist(Geo$Yc[1], Geo$Xc[1], Geo$Yc[2], Geo$Xc[2])  = 2.943368
#   Формирование матрицы расстояний (км) между точками отбора проб
library(geosphere) ; D_geo <- as.dist(distm(Geo)/1000)
# Сохраняем матрицы для использования в других видах анализа
save(Species,D_spec,Envir,D_env, Geo, D_geo,file="Fito_Dmat.RData")
#  Графическое тестирование данных – отображаются координаты и
#  рисуются круги, пропорциональные биомассы на каждого участка отбора проб,
#   и выводятся номера 20 случайных участков из 159
row.names(Geo) <- Fito_Full[,2] ; Geo2 <- Geo[sample(159,20),]
plot(Geo$Xc, Geo$Yc,asp=1,  pch=21, col="white", bg="yellow",
                            cex=5*(Fito_Full[,26])/max((Fito_Full[,26])))
text(Geo2$Xc, Geo2$Yc, row.names(Geo2), cex=0.5, col="black")
#   Тест Мантеля
mantel(D_spec, D_env)  #   «Обилие видов»  ~ «Факторы среды»
mantel(D_spec, D_geo)   #   «Обилие видов»  ~ «Пространственное расположение»
mantel(D_env, D_geo)    #   «Факторы среды»  ~ «Пространственное расположение»
mantel.partial(D_spec, D_env, D_geo)  #   Связи одновременно всх трех матриц
```

5.4. Иерархический кластерный анализ и бутстрепинг деревьев

Простота и наглядность графического восприятия делают построение деревьев классификации популярным механизмом анализа и визуализации данных. В экологии – это прекрасный способ абстрагироваться от континуальной сущности объектов и найти характерные разрывы в их непрерывности, а в биоинформатике и филогении, в основе которых лежит поиск типологии дискретных структур, анализ деревьев иерархии является ключевым методом исследований (Дюран, Оделл, 1980; Классификация…, 1980).

Кластерный анализ – нетипичный статистический метод, поскольку в нем нет проверки каких-либо гипотез. Любой алгоритм кластеризации может считаться результативным, если при использовании метрики дистанций $d(x, y)$ можно найти такое разбиение объектов на группы, что расстояния между объектами из одной группы в целом будут меньше ε, а между объектами из разных групп больше ε, где $\varepsilon > 0$ – задаваемый уровень сходства. И только пользователь может оценить качество построенных деревьев; он же и решает, являются ли эти структуры интересными и заслуживающими предметной интерпретации. Поскольку здесь на первый план неизбежно выдвигаются субъективные обстоятельства, то для реализации кластерного анализа используются самые различные математические методы. Важно, чтобы они обеспечили два наиболее важных критерия качества классифицирования: компактность групп и их разделяемость.

Наиболее популярны последовательные иерархические алгоритмы, которые отталкиваются от матрицы парных расстояний между "листьями" и строят дерево, основываясь на несложных правилах агрегирования (найти ближайшего соседа, дальнего соседа или наилучшую среднюю связь). Неиерархические итерационные процедуры, такие как метод *k*-средних, пытаются найти наилучшее разбиение, ориентируясь на некий критерий оптимизации "близости", не строя при этом полного дерева. Если удается задать

некоторую вероятностную модель процесса, то можно построить дерево, наилучшим образом воспроизводящее модель разбиения, с использованием, например, метода максимального правдоподобия. Другим современным подходом к кластеризации объектов являются алгоритмы типа нечетких *c*-средних и Гюстафсона-Кесселя, которые ищут кластеры в пространстве нечетких множеств в форме эллипсоидов, что делает их более гибкими при решении различных задач.

В условиях многообразия формул для метрик дистанций и методов кластеризации возникают естественные вопросы: какие алгоритмы следует предпочесть в конкретных условиях обработки данных и как оценить устойчивость и достоверность полученного решения. Здесь обычно ссылаются на эмпирическое правило – устойчивая типология сохраняется при изменении методов кластеризации. Таким образом, в достаточной мере оценить адекватность решения часто невозможно, не прибегая к помощи альтернативных методов. Хотя в теоретическом плане эта проблема не решена, рассмотрим раличные приемы анализа результатов кластеризации.

Вернемся в качестве примера к анализу сходства и своеобразия 13 станций наблюдения на р. Сок и его притоке р. Байтуган по данным гидробиологической съемки [П2]. Ограничимся списком из 129 наиболее представительных видов макрозообентоса и подсчитаем число проб, в которых встретился каждый вид. Выполним нормирование от 0 до 1 исходной матрицы частот и в качестве метрики дистанций рассчитаем евклидово расстояние между каждой парой станций наблюдений в многомерном пространстве видов. Осуществим кластерный анализ с использованием различных алгоритмов и рассмотрим основные задачи, стоящие при оценке полученных результатов, и приемы их решения.

1. Какая из классификаций в наибольшей степени отражает близость объектов в исходном признаковом пространстве? Для этого могут быть использованы кофенетическая корреляция (cophenetic correlation) или различные ранговые индексы.

Кофенетическая дистанция между двумя объектами, легко считываемая с полученной дендрограммы, – это уровень внутригрупповых различий, при котором объекты были впервые объединены в один кластер. Кофенетическая корреляция вычисляется как линейный коэффициент корреляции Пирсона (или ранговый коэффициент Спирмена) между всеми расстояниями, полученными по дереву иерархии, и элементами исходной матрицей дистанций.

Рассчитаем коэффициенты кофенетической корреляции ρ для четырех версий иерархической классификации представленного примера по алгоритмам "ближайшего соседа" ($\rho = 0.611$), "дальнего соседа" ($\rho = 0.781$), средней связи ($\rho = 0.796$) и методом минимальной дисперсии Варда ($\rho = 0.762$). Корреляционные зависимости удобно рассматривать в виде диаграммы Шепарда (см. рис. 5.9), на которой легко усмотреть отличия между различными вариантами кластеризации. Достоверность кофенетической корреляции легко оценить с использованием стандартного рандомизационного теста (см. раздел 2.2) или функции mantel.randtest(…), реализующей пермутацию в рамках теста Мантеля (см. раздел 5.4). Коэффициенты ρ для всех выполненных версий классификации оказались статистически значимыми при $\alpha = 0.001$.

С использованием кофенетической корреляции нельзя сравнить два разбиения, если одно из них получено неиерархическим методом. Для такого сравнения используют таблицу сопряженности кластеров, оценивающую частоты совпадений при отнесении исходных объектов к группам. Например, можно установить, что уровень совпадения разбиений с использованием алгоритма средней связи (рис. 5.9б) и метода *k*-средних после выделения 4-х кластеров составляет для нашего примера 92.3%. Поскольку метки классов сравниваемых разбиений могут быть перепутаны, то предварительно выполняется оптимизационная процедура перестановки строк и столбцов таблицы сопряженности, обеспечивающая максимум суммы элементов по главной диагонали.

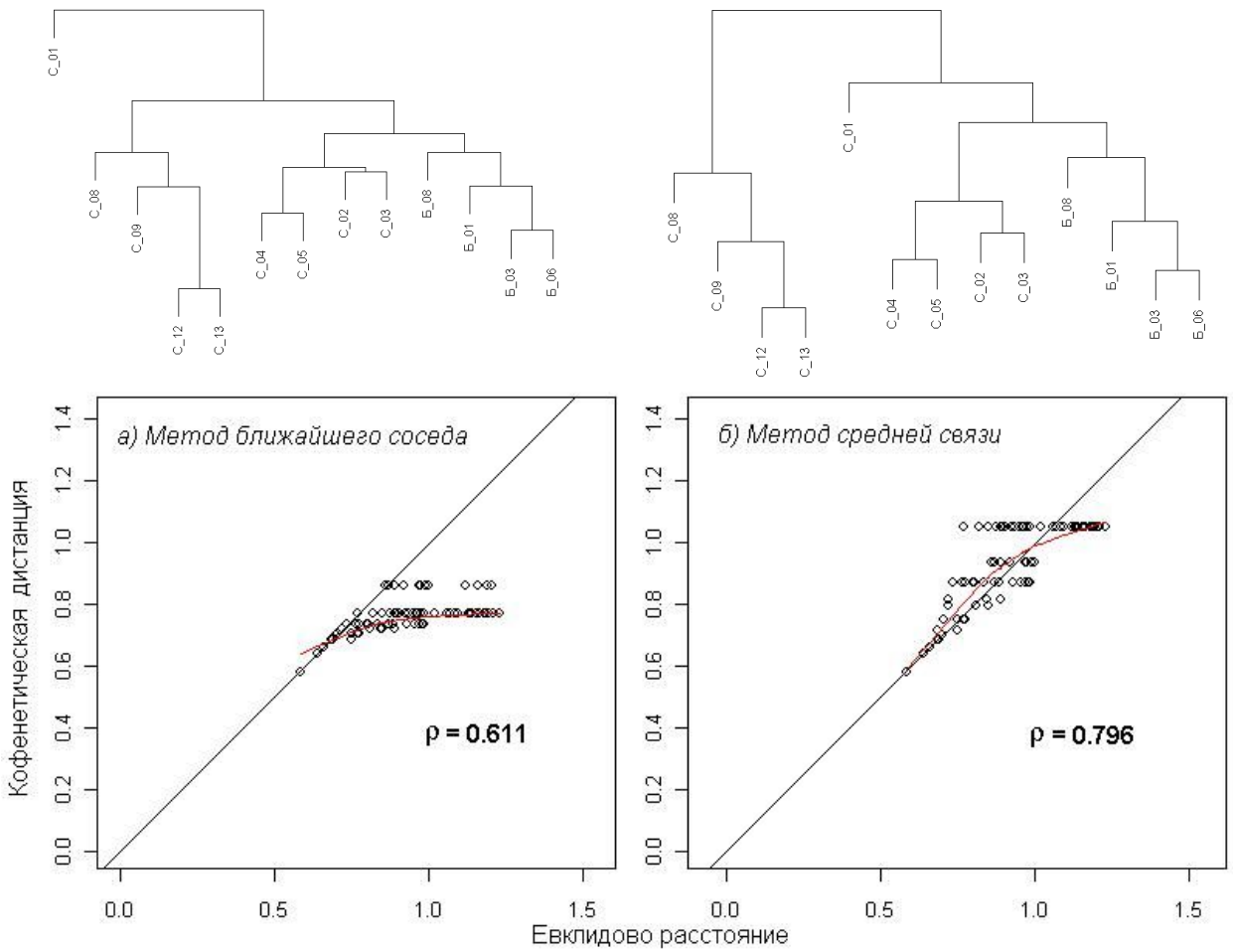
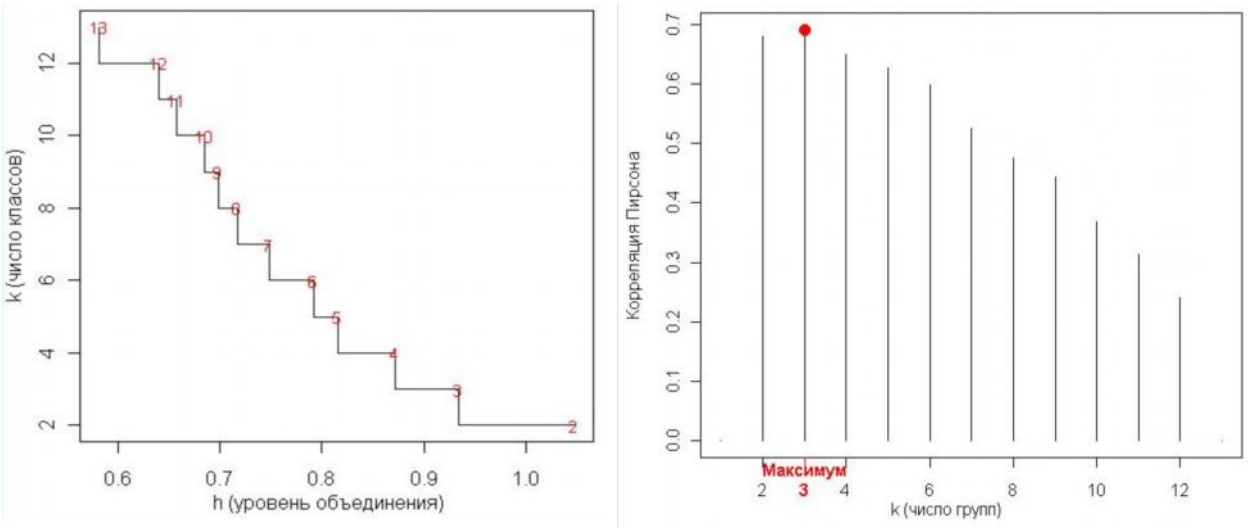


Рис. 5.9. Диаграмма Шепарда для оценки формы корреляции (ρ) между кофенетической дистанцией и евклидовым расстоянием для двух методов кластеризации – ближайшего соседа (а) и средней связи (б)

2. Как оценить количество классов наилучшего разбиения и выбрать оптимальный уровень "обрезки" дерева? Визуально рекомендуется для этого использовать различные диаграммы типа "каменной осыпи" (рис. 5.10а), изменения ширины "силуэта" или график зависимости $(\alpha_{i+1} - \bar{\alpha})/s_\alpha$ от числа кластеров, где α_i – последовательность уровней слияния групп, $\bar{\alpha}$ – среднее по первым i использованным уровням, s_α – стандартное отклонение. Другой возможный путь состоит в том, чтобы выбрать несколько подходящих разбиений, сформировать для каждой пары возможных альтернатив таблицу сопряженности и рассчитать меру связи частот совпадающих группировок (например, на основе статистики χ^2 – см. раздел 3.2).

Д. Боркард с соавторами (Borcard et al., 2011) предлагают использовать для оценки числа классов бинарную таблицу разбиений (или матрицу принадлежности объектов к группам – matrices group membership). Она представляет собой матрицу дистанций, элементы b_{ij} которой равны 1, если объекты i и j принадлежат к разным поддеревьям, и 0 в противном случае. Для выбора уровня отсечки используется следующий алгоритм:

- рассматриваются все возможные варианты "распила" дерева и для каждого из них формируются бинарные матрицы разбиений $\mathbf{B}_i$;
- рассчитываются статистики Мантеля R_i , полученные как произведения $\mathbf{B}_i$ и исходной матрицы дистанций **D**;
- оптимальным считается уровень агрегирования, доставляющий максимум R_i. (см. рис. 5.10б).



а) б)

Рис. 5.10. Диаграммы для оценки оптимального уровня обрезки дерева иерархии: (а) типа "каменной осыпи" и (б) "силуэт" статистики Мантеля

3. Насколько достоверно топология дерева отображает предупорядоченность исходных объектов? Какие фрагменты древовидной структуры являются "слабым звеном" в полученной классификационной конструкции?

Естественный способ проверить эти гипотезы состоит в том, чтобы взять повторные выборки, построить для каждой повторности свое дерево и вычислить частоту встречаемости каждого фрагмента в сформированной последовательности (Felsenstein, 1985). Разумеется, здесь невозможно обойтись без бутстрепа: на рис. 5.11 показано, как вычисляется бутстреп-вероятность BP встречаемости произвольного узла B-C. Обычно фрагменты древовидной структуры считаются достоверными, если с ветвями бутстрепного дерева связывается вероятность, превышающая 70-80% .

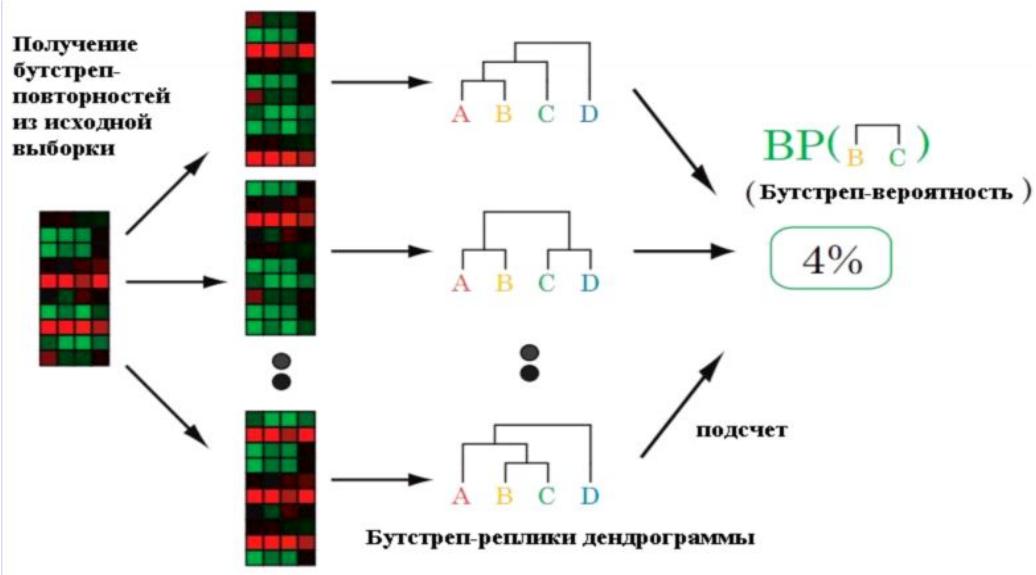


Рис. 5.11. Схема вычисления бутстреп-вероятностей фрагментов дерева классификаций

Х. Шимодара (Shimodaira, 2002), сравнивая центры распределения исходной и бутстреп-выборок, показал, что величина BP является приближенной оценкой вероятности появления узла в дереве. Несмещенную оценку AU (approximately unbiased) можно получить, выполнив повторную серию бутстрепа в различных масштабах (multiscale bootstrap resampling). Для этого отдельно вычисляют BP-значения, формируя

бутстреп-выборки разного объема: например, 0.5n, 0.6n, …, 1.4n, 1.5n, где n – объем исходной выборки. Несмещенная бутстреп-вероятность AU находится аппроксимацией ряда полученных значений BP. Наилучшие оценки AU для каждого кластера дендрограммы, найденные путем подбора параметрических моделей с использованием метода максимального правдоподобия, могут быть получены с использованием пакетов pvclust и scaleboot статистической среды R.

Дерево классификации, на которое нанесены значения AU/BP для метки каждого узла, показано на рис. 5.12.

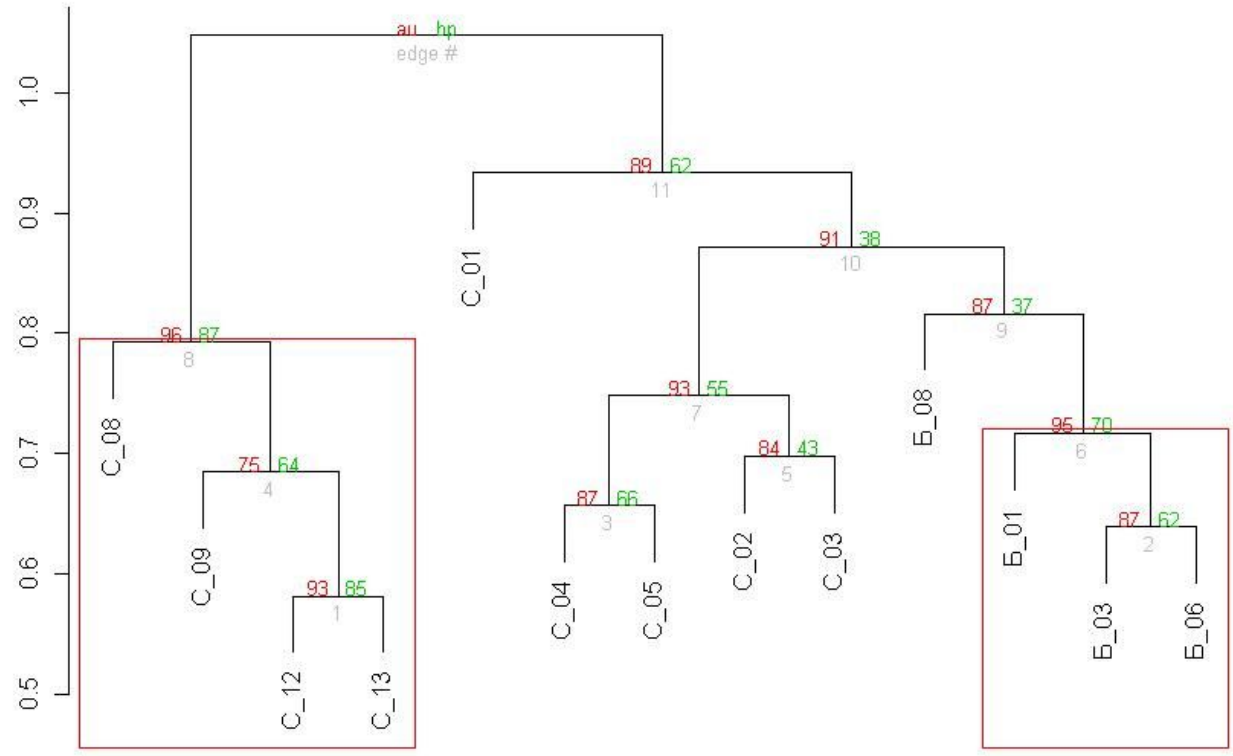


Рис. 5.12. Дерево классификации станций р. Сок-Байтуган, полученное методом средней связи по матрице нормированных евклидовых дистанций с нанесенными оценками бутстреп-вероятностей ветвей; рамками отмечены ветви с доверительной вероятностью > 95%

К разделу 5.4:

```
#  Алгоритмы оценки результатов кластерного анализа
library(xlsReadWrite)  #  Загрузка данных из файла Excel
A <- read.xls("Сок1.xls", sheet = 1, rowNames=TRUE)
library(ade4); library (vegan) ; library (cluster)
# Расчет матрицы нормированных евклидовых расстояний
A[is.na(A)] <- 0 ; A.norm <- decostand(t(A), "normalize"); spe.ch <- vegdist(A.norm, "euc")
# Построение деревьев иерархии различными методами
# Метод  одиночной  связи, или  метод «ближайшего  соседа» (Single linkage)
spe.ch.single <- hclust(spe.ch, method="single") ; plot(spe.ch.single)
# Метод  полных  связей,  или  метод «дальнего  соседа» (Complete linkage).
spe.ch.complete <- hclust(spe.ch, method="complete") ; plot(spe.ch.complete)
# Метод невзвешенного попарного  среднего (UPGMA -Unweighed pair-group  average)
spe.ch.UPGMA <- hclust(spe.ch, method="average") ; plot(spe.ch.UPGMA)
# Невзвешенный  центроидный  метод ( Unweighed pair-group centroid).
spe.ch.centroid <- hclust(spe.ch, method="centroid") ; plot(spe.ch.centroid)
# Метод  Варда минимальной дисперсии кластеризации(Ward's method).
spe.ch.ward <- hclust(spe.ch, method="ward");
spe.ch.ward$height <- sqrt(spe.ch.ward$height); plot(spe.ch.ward)
#  Cophenetic correlation - кофенетическая корреляция
spe.ch.single.coph <- cophenetic(spe.ch.single)
mantel.rtest(spe.ch, spe.ch.single.coph,nrepet = 999)
spe.ch.comp.coph <- cophenetic(spe.ch.complete)
```

```
mantel.rtest(spe.ch, spe.ch.comp.coph,nrepet = 999)
spe.ch.UPGMA.coph <- cophenetic(spe.ch.UPGMA)
mantel.rtest(spe.ch, spe.ch.UPGMA.coph,nrepet = 999)
spe.ch.ward.coph <- cophenetic(spe.ch.ward)
 mantel.rtest(spe.ch, spe.ch.ward.coph,nrepet = 999)
# Вывод диаграммы Шепарда
plot(spe.ch, spe.ch.single.coph, xlab="Евклидово расстояние",
    ylab="Кофенетическая дистанция", asp=1, xlim=c(0,sqrt(2)), ylim=c(0,sqrt(2)),
  main=c("Метод ближайшего соседа",paste("Po = ", round(cor(spe.ch, spe.ch.single.coph),3))))
abline(0,1);lines(lowess(spe.ch, spe.ch.single.coph), col="red")
#  Формирование таблицы сопряженности для разных уровней обрезки дерева
g24 <- cutree(spe.ch.UPGMA, k = 1:5) ; table(grp2=g24[,"2"], grp4=g24[,"4"])
#  Построение разбиения методом k-средних
spe.ch.kmeans <- kmeans(spe.ch, 4, nstart = 5 ) ; spe.ch.kmeans$cluster
plot(spe.ch, col = spe.ch.kmeans$cluster) ; points(spe.ch.kmeans$centers, col = 1:5, pch = 8)
#  Формирование таблицы сопряженности для сравнения разбиений
#  методом средней связи и k-средних
cont.table <- table(spe.ch.kmeans$cluster,as.vector(cutree(spe.ch.UPGMA, k = 4))) ;
print(cont.table)
## Нахождение оптимального соотношения частот двух классификаций
#  (т.е. достигнут максимум диагональных элементов)
library(e1071) ; class.match <- matchClasses(as.matrix(cont.table),method="exact")
 print(cont.table[,class.match])
#  Вывод диаграммы уровней объединения типа "каменной осыпи"
plot(spe.ch.UPGMA$height, ncol(A):2, type="S",
              ylab="k (число классов)", xlab="h (уровень объединения)", col="grey11")
text(spe.ch.UPGMA$height, ncol(A):2, ncol(A):2, col="red", cex=0.8)
# Оптимальное число кластеров, вычисляемое с использованием статистики Мантеля
# Функция вычисления бинарной матрицы распределения объектов по группам
grpdist <- function(X)   {
       gr <- as.data.frame(as.factor(X)) ;   distgr <- daisy(gr, "gower") ;   distgr   }
grpdist(cutree(spe.ch.UPGMA, 4))
# Вычисления, основанные на классификации по алгоритму средней связи
kt <- data.frame(k=1:ncol(A), r=0)
for (i in 2:(ncol(A)-1)) {  gr <- cutree(spe.ch.UPGMA, i); distgr <- grpdist(gr)
       mt <- cor(spe.ch, distgr, method="pearson") ;     kt[i,2] <- mt }
kt ; k.best <- which.max(kt$r)
# Диаграмма оптимального числа кластеров с использовнием статистики Мантеля
plot(kt$k, kt$r, type="h", xlab="k (число групп)", ylab="Корреляция Пирсона")
axis(1, k.best, paste("Максимум", k.best, sep="\n"), col="red", font=2, col.axis="red")
points(k.best, max(kt$r), pch=16, col="red", cex=1.5)
# График "Силуэт" для окончательного разбиения на кластеры
cutg <- cutree(spe.ch.UPGMA, k=4) ; sil <- silhouette(cutg, spe.ch)
silo <- sortSilhouette(sil) ; rownames(silo) <- col.names(A)[attr(silo,"iOrd")]
plot(silo,    cex.names=0.8, col=cutg+1, nmax.lab=100)
#------------------------  Бутстрепинг деревьев
#  Бутстреп деревьев и расчет смещенных и несмещенных вероятностей для узлов деревьев
library(pvclust)
cluster.bootstrap <- pvclust(t(A.norm), nboot=1000, method.dist="euclidean")
summary(cluster.bootstrap) ; plot(cluster.bootstrap) ;  pvrect(cluster.bootstrap)
#  Подбор параметрических моделей AU-вероятностей на основе максимума правдоподобия
library(scaleboot) ; fm <- sbfit(cluster.bootstrap) ; summary(fm) ; plot(fm,legend="topleft")
```

5.5. Алгоритмы оценки оптимальности разбиения на классы

Общепринятая методика оценки качества найденного разбиения на классы основана на интерпретируемости и повторяемости. Если одна и та же закономерность проявляется при использовании различных вариантов или методов классификации, отличаясь лишь в некоторых деталях, то аналитик приходит к мнению, что основная тенденция изменчивости структуры экосистемы найдена. Строгость и стройность этому

субъективному подходу могут придать количественные методы оценки надежности и статистической значимости найденных группировок.

Использование иерархических методов кластеризации матриц, содержащих большое ($n > 100$) число объектов, не всегда оправдано, т.к. визуальный анализ обширных деревьев становится затруднительным. В этом случае становится предпочтительным обратиться к математически более представительным неиерархическим алгоритмам, таким как методы *k*-средних или *k*-метоидов (*k*-means, *k*-medoids). Напомним, что в обоих случаях ищутся разделяемые между собой *центроиды*, т.е. центры сгущений точек с минимальными расстояниями внутри каждого кластера (*метоид* – это центроид, координаты которого смешены к ближайшему из исходных объектов данных). Здесь k – фиксированное число кластеров разбиения, априори задаваемое аналитиком, которое, безусловно, далеко не всегда выбирается оптимальным. Поэтому первой задачей кластерного анализа является подбор оптимального значения k, доставляющего максимум некоторому критерию, оценивающему одновременно меру однородности точек в пределах одного кластера и меру удаленности точек, принадлежащих разным кластерам.

Рассмотрим возможные критерии выбора наилучшего разбиения (goodness-of-fit) на примере [П3], представленном нами ранее в разделах 4.5 и 5.3. Матрица **D** размерностью 159×159 евклидовых расстояний между каждой парой из 159 геоботанических описаний была получена после преобразования Хеллингера значений биомассы 23 видов растений, произрастающих в дельте р. Волга. Для кластерного анализа воспользуемся процедурой поиска **"**разделяемости вокруг метоидов**"** (Partitioning Around Medoids – Kaufman, Rousseeuw, 2005), которая является наиболее устойчивой версией метода *k*-средних и эффективна для широкого диапазона различных мер расстояния. При этом максимизируется функция $F = \min \sum_{i=1}^{n} \sum_{j=1}^{n} d(i,j) z_{ij}$, где z_{ij} – бинарная матрица, описывающая различные варианты разбиения.

Для каждого найденного кластера может быть вычислена **"**ширина силуэта**"** $s_i = \frac{b(i) - a(i)}{\max[b(i), a(i)]}$, где $a(i)$ – среднее расстояние между объектами i-го кластера, $b(i)$ – среднее расстояние от объектов i-го кластера до другого кластера, самого близкого к i-му. **"**Средняя ширина силуэта**"** $\bar{s}$ (average silhouette width) определяет качество проведенной кластеризации и может явиться одним из критериев goodness-of-fit.

Для нахождения в статистической среде R оптимального k_{opt} числа разбиений на классы 159 геоботанических описаний достаточно выполнить 157 раз функцию pam(…) с изменяющимся значением параметра k от 2 до (n – 1). Наибольшее значение средней ширины силуэта $\bar{s}$ = 0.339 имело место при разбиении исходных точек на k = 11 кластеров – см. рис. 5.13.

Зададимся теперь целью оценить статистическую связь между оптимальной группировкой пробных площадок по их флористической близости и сопряженными факторами среды (увлажненностью почвы и ионным составом водной вытяжки). Проверку значимости нулевой гипотезы для каждого показателя в отдельности выполним различными способами: в процессе однофакторного дисперсионного анализа на основе теоретического распределения статистики $F(10, 148)$ и пермутационного теста (см. раздел 2.2), а также с использованием непараметрического H-критерия Краскела-Уоллеса, который оценивает сдвиги параметров положения групп, сравнивая ранги элементов вариационного ряда.

Результаты анализа в табл. 5.5 свидетельствуют о том, что разбиение площадок на 11 групп с полным основанием может быть использовано для объяснения изменчивости увлажненности и засоленности почвы.

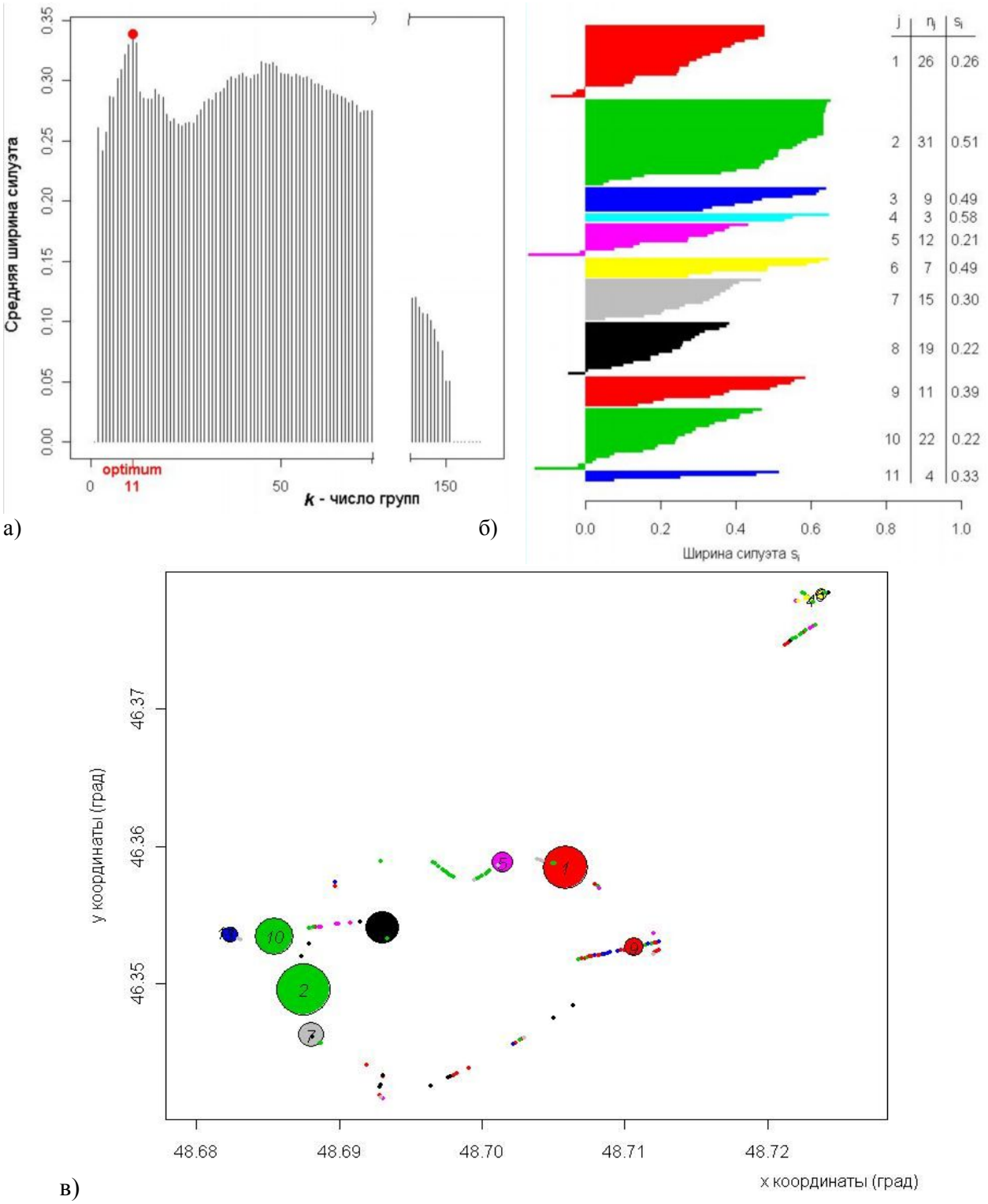


Рис. 5.13. Нахождение числа классов разбиения 159 геоботанических описаний, приводящего к максимуму средней ширины силуэта (*а*), график силуэтов кластеров оптимальной классификации (*б*) и расположение кластеров на картосхеме (*в*)

Проверка статистической значимости в целом для некоторой фиксированной группировки многомерных наблюдений, заданных матрицей расстояния, может быть выполнена с использованием двух практически идентичных непараметрических процедур: *перестановки с анализом множественного отклика* (MRPP, multi-response permutation procedure; Mielke, 1984; McCune et al., 2002) и имитационный анализ сходства ANOSIM (analysis of similarities; Clarke, 1993). Методы практически не накладывают никаких

ограничений на анализируемые данные, однако применимы только в случае простой схемы исследований, основанной на одноуровневой классификации.

Таблица 5.5. Результаты однофакторного дисперсионного анализа связи группировки пробных площадок по флористическому сходству с ионным составом почвы

Факторы	Однофакторный дисперсионный анализ			Метод Краскела-Уоллиса	
	F-критерий	p_{param}	p_{rand}	критерий χ^2	*p*-значение
Карбонаты CO_4^{2-}	2.81	0.003	0.002	25.25	0.0048
Сульфаты SO_4^{2-}	7.12	<0.0001	0.001	46.35	<0.0001
Хлориды Cl^-	9.06	<0.0001	0.001	60.12	<0.0001
Кальций $Ca2^+$	4.09	<0.0001	0.001	35.2	0.0001
Магний Mg^{2+}	7.04	<0.0001	0.001	56.16	<0.0001
Натрий Na^+	11.9	<0.0001	0.001	63.9	<0.0001
Анионы	8.66	<0.0001	0.001	57.02	<0.0001
Катионы	8.66	<0.0001	0.001	57.02	<0.0001
Минерализация	8.66	<0.0001	0.001	57.02	<0.0001
Высота	25.92	<0.0001	0.001	100.28	<0.0001

Предположим, что множество n объектов разбито на $к$ групп так, что $\sum_{i=1}^{k} n_i = n, \quad n_i \geq 2$. Процедура MRPP проверяет нулевую гипотезу, что расстояния между любой парой объектов в многомерном пространстве признаков никак не зависят от заданной классификации, т.е. внутрикластерные и межкластерные расстояния статистически эквивалентны. Суть метода заключается в следующем:

- вычисляется матрица расстояний **D** между всеми n объектами с использованием любой подходящей метрики;
- находится среднее расстояние d_i между объектами каждой выделенной группы $i = 1, ..., k$ и средневзвешенный индекс внутригруппового расстояния $\delta_{obs} = \sum_{i=1}^{k} \frac{n_i}{n} d_i$; для локальных групп, у которых $d_i > \delta_{obs}$, можно предположить ослабленную выраженность группировки;
- запускается процедура случайной перестановки между группами строк и столбцов матрицы **D** и каждый раз вычисляются средние значения различий объектов δ_{sim} внутри перемешанных групп;
- процедура рандомизации повторяется достаточно большое число раз, чтобы можно было построить статистическое распределение величины δ;
- подсчитывается число случаев, когда значение δ_{sim} не превышает эмпирическое значение δ_{obs}, и вероятность p этого события определяет вероятность ошибки 1-го рода при проверке нулевой гипотезы (это соответствует площади под кривой распределения δ левее значения δ_{obs}).

Кроме значения *p,* необходимого для проверки статистической значимости группировки, может быть вычислена также оценка эффективности априорной классификации. Индекс внутригрупповой однородности, независимый от объема выборки и показывающий превышение среднего уровня сходства объектов внутри классов по сравнению со случайным размещением, может быть вычислен как $A = 1 - \delta_{obs} / m_{\delta}$, где m_{δ} - математическое ожидание δ при справедливости нулевой гипотезы. Если $A < 0$, то однородность внутри выделенных групп ниже, чем при случайном их формировании, т.е при A = 1 группы состоят из совершенно идентичных объектов, а при A > 0.2 можно говорить о существенной ценности классификации.

Для оценки межгрупповых различий в другом методе ANOSIM вычисляется статистика R, являющаяся некоторой модификацией коэффициента корреляции. Она

всегда находится в интервале [-1; +1] и основана на разности средних рангов различия объектов между группами b и различия внутри групп w: $R = \frac{4(b-w)}{n(n-1)/4}$, где n – общее число анализируемых объектов. Отрицательное значение R свидетельствуют о том, что внутренняя неоднородность сформированных групп настолько велика, что превышает межгрупповые отличия, а чем ближе ее значение к +1, тем больше различаются сформированные группы. Статистическая значимость эмпирического значения статистики R оценивается с помощью описанного выше перестановочного теста.

Применение методов MRPP и ANOSIM в представленном примере дало основания сделать следующие выводы (см. табл. 5.6):

° поскольку качество разбиения основывается только на внутригрупповых расстояниях, оцениваемые статистики A и R почти всегда монотонно возрастают с увеличением числа классов, в связи с чем они не могут быть использованы для поиска оптимальной группировки;

° методы излишне либеральны к отклонению нулевой гипотезы и найденное разбиение практически всегда оказывается статистически значимым.

Таблица 5.6. Оценка качества разбиений пробных площадок в дельте р. Волга с использованием методов MRPP и ANOSIM

Показатель	Количество классов разбиения				
	4	11	18	40	60
Внутригрупповое расстояние δ_{obs} (δ_{sim} = 1.136)	0.84	0.605	0.537	0.400	0.327
Внутригрупповая однородность A	0.261	0.467	0.618	0.648	0.712
R-статистика ANOSIM	0.734	0.934	0.938	0.974	0.983
P-значение (MRPP и ANOSIM)	0.001	0.001	0.001	0.001	0.001

В дополнение к этой теме следует отметить, что если предположить, что данные, составляющие каждый кластер, являются нормально распределенными, то можно использовать методы разделения смесей (EM-алгоритм), а выбор оптимального числа классов осуществлять на основе AIC-критерия.

К разделу 5.5:

```
# Алгоритмы оценки оптимальности разбиения на классы
# Загрузка данных из ранее сохраненного файла (см. скрипт к разделу 5.3)
load(file="Fito_Dmat.RData") ; ls() ; library(vegan) ; library (cluster)
# Используем объекты: D_spec - матрица евклидовых расстояний между описаниями
# Geo - географические координаты пробных площадок, Envir - ионный состав почвы
# Вычисляем наилучшее разбиение методом Partitioning around medoids (PAM)
asw <- numeric(nrow(Species))
for (k in 2:(nrow(Species)-1)) ;   asw[k] <- pam(D_spec, k, diss=TRUE)$silinfo$avg.width
k.best <- which.max(asw) ; cat("", "Оптимальное число кластеров k =", k.best, "\n",
                          "со средней шириной силуэтов S=", max(asw), "\n")
plot(1:nrow(Species), asw, type="h", xlab="k (число групп)", ylab="Средняя ширина силуэтов")
axis(1, k.best, paste("optimum",k.best,sep="\n"), col="red", font=2,col.axis="red")
points(k.best, max(asw), pch=16, col="red", cex=1.5) # График изменения статистики S
spe.ch.kmeans <- pam(D_spec, k=k.best, diss=TRUE); str(spe.ch.kmeans)
# График ширины силуэтов по кластерам для наилучшего числа разбиений
plot(silhouette(spe.ch.kmeans),  cex.names=0.8, col=spe.ch.kmeans$silinfo$widths+1)
# График местоположения метоидов на картосхеме
KM <- spe.ch.kmeans$medoids ; countKM <- spe.ch.kmeans$clusinfo[,1] ;
grKM <- spe.ch.kmeans$cluster
plot(Geo$Xc, Geo$Yc, asp=1, type="n", xlab="x координаты (град)", ylab="y координаты (град)")
for (i in 1:k.best) {  cex_var = 7*(countKM[i])/max(countKM)
    points(Geo[grKM==i,2], Geo[grKM==i,1], pch=21, cex=0.5, col=i+1, bg=i+1)
    points(Geo[KM[i],2], Geo[KM[i],1], pch=21, cex = cex_var, bg=i+1, col="black")
```

```
    text(Geo[KM[i],2], Geo[KM[i],1], i, cex=1, font=3, col="black")    }
 legend("bottomright", paste("Группа ",1:k.best), pch=21,
        col=2:(k.best+1),pt.bg=2:(k.best+1),  pt.cex=2, bty="n")
# Однофакторый дисперсионный анализ (параметрический и Крускала-Уоллиса)
#   связи группировки на k.best классов с изменчивостью факторов окружающей среды
#  Используем скрипт P. Legendre - http://www.bio.umontreal.ca/legendre/indexEn.html
source("anova.1way.R") ; Fac_clu <- as.factor(spe.ch.kmeans$cluster)
attach(Envir) ; La <- colnames(Envir); ANOVA_RES <-as.vector(rep(NA,length(La)))
for (i in 1:ncol(Envir))  {  KW <- kruskal.test(Envir[,i] ~ Fac_clu)
       ANOVA_RES[i] <-list(c(Envir=La[i], anova.1way(Envir[,i]~Fac_clu, nperm=999),
       Kruskal_Wallis =paste("chi squared = ", KW$statistic,"  p.value=", KW$p.value)))  }
#  ---------------------------------
#  Функция оценки статистической значимости кластеризации (ANOSIM и MRPP)
Sign_clus <- function (k) { kmeans <- pam(D_spec, k, diss=TRUE);
     anoR <- anosim(D_spec, kmeans$cluster); mrppA <- mrpp(D_spec, kmeans$cluster);
  return (cat("Кластеров k =", k," ANOSIM  R =", anoR$statistic, " p =", anoR$signif, "\n",
       " MRPP  A =", mrppA$A,  "Дельта(эмп)=", mrppA$delta, "Дельта(ранд)=", mrppA$E.delta,
       " p =", mrppA$Pvalue, "\n"))  }
Sign_clus(4) ; Sign_clus(11) ; Sign_clus(18);  Sign_clus(40) ; Sign_clus(60)
```

5.6. Использование нечетких множеств для классификации и оценки силы связи

Традиционные принципы анализа структуры изучаемых систем предполагают, что выделяемые классы представляют собой детерминированные совокупности, т. е. каждый объект может принадлежать только к одному таксону. Ограниченность такого подхода часто приводит к аналитической неопределенности и множественности выводов при сравнении сообществ. Разумной альтернативой понятию абсолютной дискретности в классической таксономии является интерпретация компонентов систем как нечетких объектов в составе гибко настраиваемых ординационных структур.

Операции с нечеткими множествами появились, как эффективная практическая мера преодоления правила несовместимости Неймана-Заде: «повышение точности описания сложной системы становится несовместимым со здравым смыслом, поскольку сложность модели становится соизмеримой со сложностью самого объекта». В конце 80-х годов, после бурного развития технических устройств на базе нечетких контроллеров и экспертных систем, теория нечетких множеств (fuzzy sets) и нечеткая логика (fuzzy logic) становятся важными обобщениями классических математических теорий и неотъемлемой составной частью современных систем искусственного интеллекта.

Основные понятия fuzzy-концепций, впервые предложенные американским ученым Лотфи Заде (Zadeh, 1965; Bezdek, 1987), сводятся к следующему:

° множество **C** является нечетким, если существует функция принадлежности (membership function) $\mu_C(x)$, принимающая на этом множестве значения в интервале [0, 1];

° функция $\mu_C(x)$ конструируется на основе экспертных заключений или любого подходящего формального метода и оценивает степень сродства (grade) анализируемого объекта x относительно произвольного множества **C**: $\mu_C(x) = 0$ означает полную несовместимость, т. е. $x \notin \mathbf{C}$, а $\mu_C(x) = 1$ – полную принадлежность или $x \in \mathbf{C}$;

° нечеткое множество **C** задается множеством упорядоченных пар типа $C = \{x, \mu_C(x)\}$; в частном случае, если функция принадлежности принимает значение только 0 или 1, то **C** становится "четким" или обычным множеством.

Рассмотрим две задачи, решаемые с использованием аппарата нечетких множеств.

Задача классификации. Отличие принципов четкой и нечеткой классификации можно проследить на примере алгоритма *k*-средних. Пусть $D_{ir} = \sum_j (x_{ij} - v_{rj})^2$ - расстояние между каждым *i*-м объектом классификации (i = 1, 2, n), описанным набором признаков x_{ij}, и центрами тяжести v_{rj} каждого выделенного кластера из k (r = 1,2,…, k). Тогда в

общем случае кластеризацию объектов **X** можно сформулировать как следующую задачу оптимизации: найти матрицу **μ**, которая доставляла бы минимум значению критерия:

$$F_{km}(\boldsymbol{\mu}) = \sum_{i=1}^{n}\sum_{r=1}^{k}\mu_{ir}^{m}D_{ir}$$

В случае четкой кластеризации каждый объект $\mathbf{x}_i$ может принадлежать только к одному классу r, и тогда $\mu_{ir} = 1$, а остальные компоненты матрицы **μ** равны 0. Дискретный характер четкого разбиения обуславливает негладкость целевой функции, что усложняет нахождение оптимальной кластеризации. При использовании методов нечеткой таксономии функция $\mu_{ir}(\mathbf{x})$ задает в масштабе от 0 до 1 степень принадлежности каждого объекта $\mathbf{x}_i$ к каждому выделяемому классу r. Конкретные значения матрицы принадлежности **μ**, приводящие к минимуму F, могут быть найдены нелинейной оптимизацией, например, методами неопределенных множителей Лагранжа.

Экспоненциальный вес (m) в алгоритме нечетких k-средних задает уровень нечеткости получаемых кластеров: чем больше m, тем нечеткое разбиение более "размазано". "Штатный" диапазон варьирования m – от 1.2 до 2. Другим важным параметром является количество классов k, которое приходится принимать из априорных представлений о структуре данных.

Результат нечеткой классификации 13 станций р. Сок-Байтуган (см. пример [П2] к разделу 5.4) при $m = 1.5$ и $k = 3$ представлен на рис. 5.14. Каждому объекту ставится в соответствие по три меры принадлежности к классам. Например, для специфичной по составу донных животных станции 1 на р. Сок значения μ = {0.42, 0.41, 0.16} , т.е. практически с равным основанием этот фрагмент водотока может быть причислен к любому из первых двух классов (он образует самостоятельный кластер, если принять $k = 4$ – см. также для сравнения рис. 5.12).

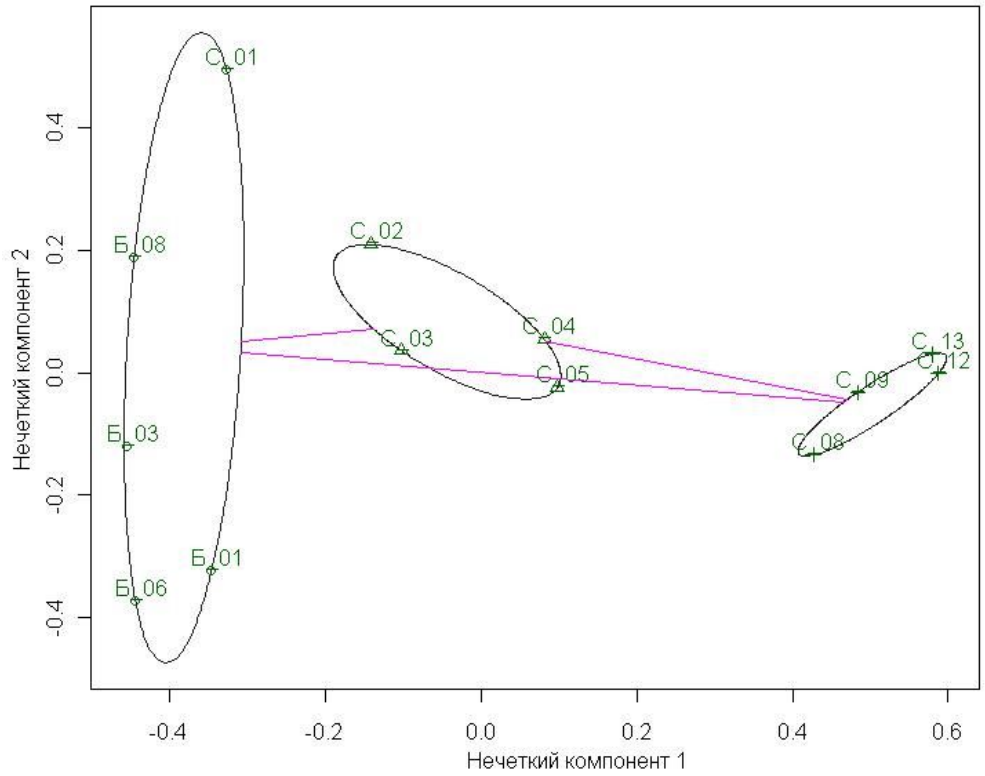


Рис. 5.14. Группировка станций р. Сок-Байтуган с использованием алгоритма нечеткой классификации

Для оценки меры нечеткости полученной классификации используется коэффициент разделения Данна (Dunn): $F_k = \sum_{i=1}^{n}\sum_{r=1}^{k}\mu_{ir}^{2}/k$, который принимает минимальное значение при полной нечеткости разбиения, когда расстояния от каждого

объекта до центра тяжести любого кластера равновелики $\mu = 1/k$. Напротив, в случае четкой кластеризации ($\mu = 1$ или 0) коэффициент Данна F_k принимает значение 1. Для представленного примера $F_k = 0.47$, а его нормированная версия, изменяющаяся от 0 до 1 и характеризующая степень нечеткости $F_k^{'} = (kF_k - 1)/(k - 1) = 0.207$.

Оценка связи видовой структуры с факторами среды. Операции с нечеткими множествами были реализованы в методе ординации сообществ, известном под аббревиатурой FSO (Fuzzy Sets Ordination – Roberts, 1986, 2008). В одномерном FSO исследователем задается один ключевой параметр среды *z*, который потенциально может оказаться ведущим градиентом, а свойства экосистемы описываются рефлексивной и симметричной матрицей любых мер подобия *d*(*x, y*) между каждой парой экологических объектов в пространстве видов.

В отличие от статистических методов оценки силы влияния фактора *z*, рассмотренных в предыдущих главах и оперирующих с реальными данными, метод FSO выполняет их предварительную *фаззификацию*, т.е. перевод в нечеткое топологическое пространство. После этого связи между компонентами экосистемы ищутся с учетом нечетких отношений между ними, задаваемых функциями принадлежности $\mu_P(x)$.

Пусть *x* и *y* – анализируемые экологические объекты (например, местообитания), а *d*(*x, y*) – сходство между этими объектами в соответствии с выбранной метрикой, такой как мера Жаккара, Съеренсена, Горна, Юла и др. Выполняется формирование четырех нечетких множеств, **P** = {*x*, $\mu_P(x)$}, **P** = **E**, **C**, **D**, **A** , где $\mu_P(x)$, – функция принадлежности, задающая на интервале [0, 1] меру близости любой пары экологических объектов из множества **P** и вычисляемая следующим образом:

° функция принадлежности $\mu_A(x)$ является набором нормированных значений фактора среды z_x для всех классифицируемых объектов *x*: $\mu_{\mathbf{A}}(x) = \dfrac{z_x - \min(z)}{\max(z) - \min(z)}$;

° две функции связывают $\mu_A(x)$ с *d*(*x, y*) и соответствуют высоким и низким значениям фактора среды *z*: $\mu_{\mathbf{C}}(x) = \dfrac{\sum_{y \neq x} d(xy)\mu_{\mathbf{A}}(x)}{\sum_{y \neq x} \mu_{\mathbf{A}}(y)}$; $\mu_{\mathbf{D}}(x) = \dfrac{\sum_{y \neq x} d(xy)[1 - \mu_{\mathbf{A}}(x)]}{\sum_{y \neq x} [1 - \mu_{\mathbf{A}}(y)]}$;

° результирующее нечеткое множество **E**, которое обычно используется для отображения на ординационной диаграмме, вычисляется с использованием оператора *антикоммутативной разности*, оценивающего контраст двух полярных нечетких множеств **C** и **D**: $\mu_E(x) = \{1 + [1 - \mu_D(x)]^2 - [1 - \mu_C(x)]^2\}/2$.

Практически метод FSO моделирует распределение оптимумов относительной таксономической насыщенности разных местообитаний на шкале градиента фактора среды, т.е. оценивается роль, которую играет переменная *z* в формировании или ограничении отдельных видов в структуре сообществ. При этом функцию принадлежности $\mu_E(x)$ можно рассматривать как одну из версий количественного представления мультивидового отклика на изменчивость фактора среды *z* при выполнении "калибровки" (Jongman et al., 1987, глава 4).

Выполним расчет компонентов нечетких множеств на примере данных о встречаемости 214 видов макрозообентоса на 13 участках речной экосистемы Байтуган-Сок. Для расчета матрицы расстояний будем использовать индекс Горна – см. формулу в разделе 5.1, которая дает наиболее стабильные результаты для количественных признаков (Boyce, Ellison, 2001).

Простым критерием значимости полученных осей нечетких множеств может служить корреляция между функцией принадлежности $\mu_E(x)$ и значениями соответствующей экологической переменной *z*. Если в распоряжении исследователя оказывается целый набор потенциальных параметров среды, определяющих

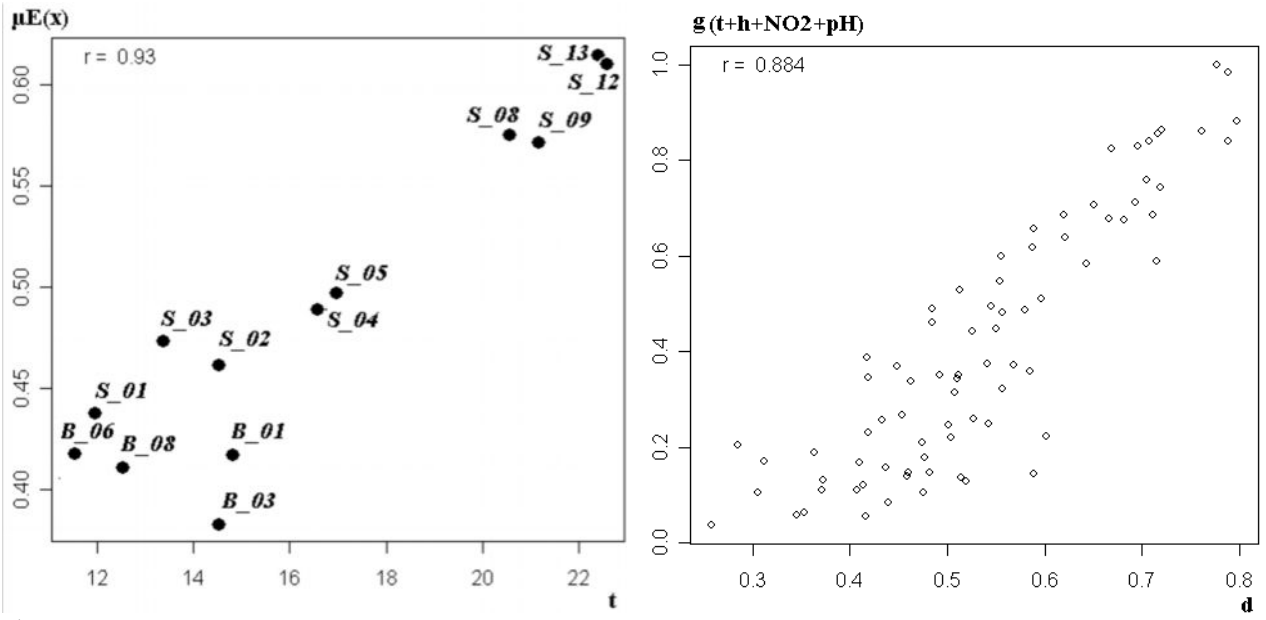
закономерность распределения популяционной плотности видов по местообитаниям, то можно последовательно провести их сравнительный анализ и факторы с высокой корреляцией считать наиболее влияющими.

Оценка *p*-значений статистической значимости коэффициентов корреляции Пирсона *r* может быть вычислена, как мы это делали в разделе 3.1, обычным параметрическим путем и с использованием рандомизации. Расчеты показали (табл. 5.7), что можно принять статистически значимым влияние 8 из 12 анализируемых факторов среды на таксономическую изменчивость донных сообществ.

Таблица 5.7. Коэффициенты корреляции Пирсона *r* между переменными среды и функцией принадлежности нечетких множеств $\mu_E(x)$; p_{norm} и p_{rand} – оценки статистической значимости *r*, полученные методом нормального приближения и рандомизацией

Переменные среды	*r*	p_{norm}	P_{rand}
Температура воды (t)	0,929	$4{,}18\cdot10^{-6}$	**0,001**
Глубина в местах отбора проб (h)	0,921	$7{,}55\cdot10^{-6}$	**0,001**
Содержание нитритного азота (N-NO2)	0,828	$4{,}72\cdot10^{-4}$	**0,001**
pH	0,822	$5{,}61\cdot10^{-4}$	**0,001**
Каменистость грунта (Stone)	0,806	$8{,}67\cdot10^{-4}$	**0,001**
Высота над уровнем моря (H)	0,797	0,001	**0,001**
Заиленность грунта (Mud)	0,778	0,0017	**0,001**
P min – содержание минерального фосфора	0,641	0,018	**0,007**
Скорость течения (v)	0,412	0,16	0,077
Содержание кислорода (O2)	0,180	0,55	0,226
Площадь водосбора (*F*)	0,044	0,88	0,323
Бихроматная окисляемость (BO)	-0,526	0,064	0,784

Методы нечеткой логики на современном этапе не предоставляют таких разносторонних способов графической интерпретации, как классические алгоритмы ординации (многомерное шкалирование, CCA и др.). Обычно исследователю предлагается проанализировать корреляционные поля зависимостей $\mu_E(x)$ от *z* (рис. 5.15а), матрицы расстояний *d* от *z* (рис. 5.15б) или различные парные взаимодействия двух нечетких множеств.



а) *б*)

Рис. 5.15. Корреляционная связь функции принадлежности нечеткого множества $\mu_E(x)$ с температурой воды t (*а*) и комплексного ординационного расстояния $g(t+h+NO_2+pH)$ с компонентами матрицы таксономических расстояний d (*б*)

Одномерный FSO позволяет, подобно одномерному регрессионному анализу, оценить статистическую зависимость функций принадлежности размытых множеств только от одного фактора среды. Для анализа всего комплекса переменных в целом можно воспользоваться *многомерной* версией ординации нечетких множеств MFSO (Roberts, 2008). В методе MFSO для определения нечетких множеств в многомерном пространстве сначала выполняется нахождение ординационной оси относительно экологической переменной, которая в наибольшей мере объясняет вариацию композиций видов (т.е. имеет наибольший коэффициент корреляции). Вторая ось, связанная со следующим фактором, рассчитывается как перпендикуляр к первой с использованием процедуры ортогонализации Грамма-Шмидта, а ее значения соответствуют остаткам регрессии на первую ось. Таким образом, значения второго градиента нечеткого множества оценивают только ту долю влияния второй переменной, которая не может быть объяснена первым фактором. Оси третьего и высшего порядка рассчитываются аналогично и отражают долю таксономической вариации, не связанную со всеми предыдущими осями нечеткого множества.

Поскольку ортогонализация Грамма-Шмидта неэффективна при сильной взаимной корреляции анализируемых факторов, многомерный анализ выполняют после проведения шаговой процедуры FSO, которая может оказаться полезной для быстрой селекции значимых осей ординации. Шаговый алгоритм стартует с фактора, имеющего наибольший коэффициент корреляции $r(d, z)$, и последовательно включает в модель остальные переменные, оценивая их по приращению, вносимому ими в результирующее значение r. Поскольку в нашем примере некоторые переменные сильно коррелируют между собой, процедура последовательной селекции позволила исключить из дальнейшего рассмотрения три фактора среды, использование которых снижает общий коэффициент корреляции – см. результаты анализа в табл. 5.8 слева.

Таблица 5.8. Результаты шаговой процедуры селекции признаков и многомерного анализа нечетких множеств с выделением значимых осей ординации с использованием коэффициентов корреляции r; p – оценки статистической значимости приращений r, найденные рандомизацией

Переменные среды	Шаговая процедура			Многомерная процедура		
	$r(d, z)$	Прира-щение	p	$r(\mu_E(x), z)$	Прира-щение	p
Температура воды (t)	0,879			0,929		**0,001**
pH	0,8824	0,0034	0,07	0,9328	0,0038	**0,017**
Содержание нитритного азота (N-NO2)	0,8855	0,0031	0,14	*0,9337*	*0,00095*	*0,055*
Высота над уровнем моря (H)	0,8857	0,00017	0,17		*-0,002*	
Глубина в местах отбора проб (h)	0,8859	0,00024	0,18	*0,9339*	*0,0002*	*0,071*
P min – содержание мин. фосфора	*0,8859*	*-0,00002*	*0,34*			
Каменистость грунта (Stone)	*0,8839*	*-0,002*	*0,23*			
Заиленность грунта (Mud)	*0,8831*	*-0,00079*	*0,25*			

Примечание: курсивом отмечены факторы, исключаемые из рассмотрения на каждом этапе

Применение многомерной процедуры анализа нечетких множеств MFSO к оставшимся признакам (табл. 5.8 справа) позволяет выделить две относительно некоррелированные и статистически значимые оси ординации донных сообществ, основанные на температуре воды и pH.

К разделу 5.6:

```
#  Использование нечетких множеств
#  -------------  Функция нечеткой кластеризации fanny. Подробности см.
#  в обзоре В. К. Солондаева http://cafedra.narod.ru/solondaev/solond-R-fanny.pdf
#  Для нашего примера используем матрицы  A.norm и spe.ch, вычисленные в разделе 5.4
A <- read.xls("Сок1.xls", sheet = 1, rowNames=TRUE) ; A[is.na(A)] <- 0 ; library (vegan)
 A.norm <- decostand(t(A), "normalize"); spe.ch <- vegdist(A.norm, "euc")
```

```
#  Кластеризация с использованием метода нечетких c-средних и исходной таблицы
Z1.f <- fanny(A.norm, 3, metric = "SqEuclidean", diss=FALSE) ; clusplot(Z1.f)
#  Кластеризация с использованием матрицы дистанций, полученной в разделе 5.4
Z2.f <- fanny(spe.ch, 3, memb.exp = 1.5) ; clusplot(fanny(Z2.f,labels= 3, col.clus="gray11")
#  ------------  Функция нечеткой ординации  fso (…) . Подробности см.
# на сайте С.Робертса  http://ecology.msu.montana.edu/labdsv/R/labs/lab11/lab11.html
#  Загрузка данных о встречаемости видов на станциях водотока и параметрах среды
 veg <- read.table("bs_ben.txt",header=T, sep="\t")
 site <- read.table("bs_site.txt",header=T,sep="\t") ;  attach(site)
 source("abundsim.R")  # Подгружаем скрипт с функциями расчета мер подобия
 sim <- sim.abund(veg,method=2)  ;  dis.ho <- 1 - sim
 grads.fso <- fso(~~hg+Ss+v+h+O2+P_min+NO2+BO+pH+KG+IP+t,dis.ho,permute=1000)
 summary(grads.fso)  #   Результаты одномерного FSO
grads.t <- fso(t,dis.ho) ; plot(grads.t) ;  plot(grads.t,dis.ho)
#   Шаговая процедура  FSO
step.mfso(dis.ho,start=data.frame(t),add=data.frame(pH,h,hg,NO2,KG,IP))
#   Многомерная процедура  FSO
grads.mfso <- mfso(~~t+h+NO2+pH,dis.ho, scaling=2,permute=1000) ;  summary(grads.mfso)
```

5.7. Дендрограммы и оценка функционального разнообразия

Ранее в разделе 2.8 нами обсуждалось понятие функционального разнообразия сообщества, как ширины статистического интервала характерных значений (trait values) потенциально важного функционального признака *T*, который может быть оценен для каждого *i*-го вида из *s*. Для многомерного случая, когда ищется функциональное разнообразие композиции видов по совокупности характерных признаков, была разработана следующая процедура (Petchey, Gaston, 2002, 2007), основанная на анализе иерархического дерева классификации:

- исходными данными является матрица из характерных значений *m* различных функциональных признаков T_{ij}, установленных для каждого *i*-го вида расчетным путем или по эмпирическим данным, $i = 1, 2, ..., s; j = 1, 2, ..., m$;
- рассчитывается матрица дистанций **D** размерностью $s \times s$ между каждой парой видов в пространстве их характерностей с использованием, например, евклидового расстояния или метрики Гувера;
- с использованием любого подходящего алгоритма, например, средней связи, строится полное дерево классификации видов;
- формируется вектор **b**, включающий длины *l* всех ветвей дерева, и матрица инцидентности, определяющая, какая ветвь принадлежит каждому виду из *s*;
- функциональное разнообразие FD любой композиции из *s* определяется как сумма длин всех ветвей дерева, принадлежащих анализируемому подмножеству видов.

Обоснование предлагаемой процедуры вполне прозрачно. Действительно, чем больше различие между видами в многомерном пространстве признаков, тем больше фенотипическое разнообразие сообщества (насколько тождественны между собой функциональное и фенотипическое разнообразие, оставим за рамками наших рассуждений). Представляется также обоснованным оценивать дисперсию многомерных значений характерностей как сумму длин ветвей дерева. Например, очевидно, что FD композиции видов S1-S6 на рис. 5.16 существенно превышает разнообразие для S7-S10.

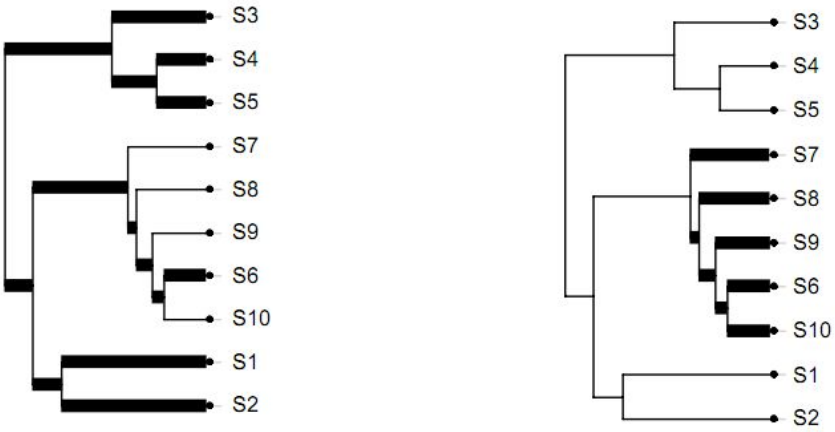
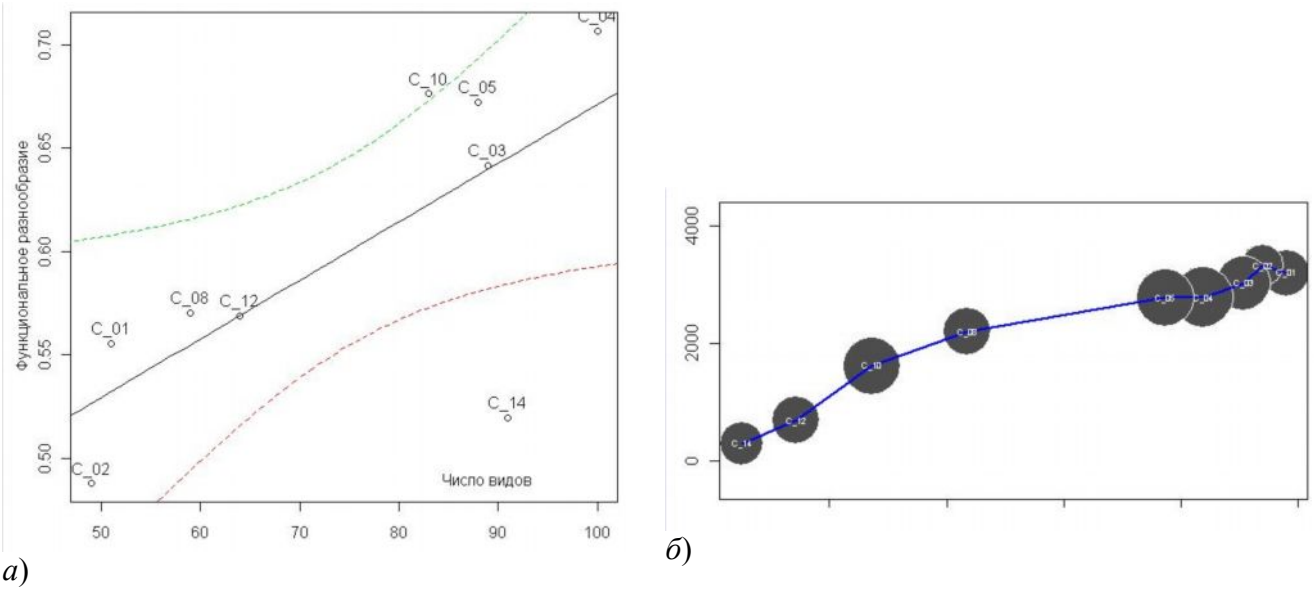


Рис. 5.16. Сравнение функционального разнообразия двух сообществ, оцениваемое как сумма длин ветвей, показанных жирными линиями

Нам не удалось подобрать необходимое количество фенотипических признаков, характерных для донных организмов. Поэтому введем понятие факториального разнообразия RD, оценивающего степень дивергенции между видами, сосуществующими в сообществе, по отношению к многомерному комплексу внешних факторов. Элементы матрицы **T** размерностью 277×12 рассчитаем как средневзвешенные значения параметров окружающей среды (m = 12 – см. табл. 5.7) для каждого из видов макрозообентоса (s = 277), встретившихся в р. Сок: $T_{ij} = \sum_{k=1}^{9} t_{ijk} a_{ik}$; $\sum_{k=1}^{9} a_{ik} = 1$, k = 1, 2, ..., 9 – последовательность станций, на которых выполнялся комплекс наблюдений.

По результатам расчетов установлена статистически значимая корреляция факториального разнообразия от числа видов (r = 0.71, p = 0.032 – см. рис. 5.17а), что является обычным явлением в случае функционального разнообразия, но требует дополнительного осмысления для нашего примера. Более прост для интерпретации тот факт, что сообщества в истоках (С-1 и С-2) и в устьевом участке С-14 имеет более компактную локализацию на дереве и, соответственно, узкий диапазон толерантности к комплексу факторов среды, чем на участках среднего течения реки.



а) *б*)

Рис. 5.17. Регрессионная связь функционального разнообразия FD с чисом видов (*а*) и распределение FD по профилю р. Сок в географических координатах (*б*)

К разделу 5.7:

```
##  Расчет функционального разнообразия по  (Petchey, Gaston 2002)
##  Подробности см.  http://www.ieu.uzh.ch/petchey/Code/code.html
source("Xtree.r")   ##   Загрузка скрипта функций, вычисляющих длину ветвей дерева
##   ---------------- Выполнение расчетов
library(xlsReadWrite); library(vegan)
SP <- read.xls("Сок Функц.xls",sheet = 3, rowNames=TRUE)  #  Частоты встречаемости видов
TS <- read.xls("Сок Функц.xls",sheet = 2, rowNames=TRUE)  #  Факторы среды на участках
#   Расчет толерантности каждого вида к каждому из факторов
TS_norm <- decostand(TS,"standardize")
species.traits <- as.matrix(SP)%*%as.matrix(TS_norm)
#  Загрузка списков видов для каждого участка
community.composition <- read.xls("Сок Функц.xls",sheet = 1, rowNames=FALSE)
Distance.method <- "euclidean" ; Cluster.method <- "average"
distances <- dist(species.traits)
tree <- hclust(distances) ; xtree <- Xtree(tree)
FD_2 <- Getlength(xtree, community.composition)
rownames(FD_2)<- c("C_01","C_02","C_03","C_04","C_05","C_08","C_10","C_12","C_14")
#  Построение графика линейной регрессии с доверительными интервалами
attach(FD_2); fit <- lm(FD.new ~ S); summary(fit) ; x <- seq(0.9*min(S), 1.1*max(S), len=100)
pre <- predict(fit, data.frame(S=x), interval="confidence")
plot(FD_2,xlab="Число видов", ylab="Функциональное разнообразие")
text(FD_2, row.names(FD_2), cex=1, pos=3) ; matplot(x, pre, type="l", lty=c(1,2,2), add=TRUE)
#  Построение графика распределения FD по профилю реки в относительных геокоординатах
TTB.Coor <- data.frame(c(9789,9397,9068,8385,7734,4347,2715,1429,500),
              c(3207,3324,3023,2778,2777,2193,1628,694,300))
 colnames(TTB.Coor)<- c("X","Y")
plot(TTB.Coor, asp=1,  pch=21, col="white", bg="grey30", cex=8* FD.new/max(FD.new))
lines(TTB.Coor, lwd=2, col="blue") ; text(TTB.Coor, row.names(FD_2), cex=0.5, col="white")
```

6. КЛАССИФИКАЦИЯ, РАСПОЗНАВАНИЕ И СНИЖЕНИЕ РАЗМЕРНОСТИ

6.1. Методы многомерной классификации и ординации

"Операционным полем" аналитических действий в области экологии сообществ являются наборы данных, организованные в виде трех таблиц: матрицы **Y** ($n \times m$) показателей популяционной плотности m видов, зарегистрированных в ходе наблюдений на подмножестве n местообитаний, матрицы **X** ($n \times q$), содержащей измерения совокупности q факторов среды в каждой точке взятия экологических проб и матрицы **B** ($m \times p$), содержащей набор p аутэкологических характеристик изучаемых видов (таксонов). Важнейшими задачами при этом являются: а) выделение эколого-ценотических и функциональных групп видов или б) группировка местообитаний (районирование).

В общем случае таблицы с результатами наблюдений могут быть геометрически интерпретированы как существенно "размытые" сгущения точек (объектов) в многомерном пространстве признаков. При этом важнейшими задачами статистического анализа являются ординация и классификация. *Ординация* основывается на принципе "континуальности" (непрерывности) и ищет упорядоченную последовательность проекций изучаемых объектов на главные оси пространства, с которыми потенциально может быть связана интерпретация научных гипотез. *Классификация*, напротив, исходит из принципа "дискретности" (или разделяемости) и выполняет статистический анализ результатов разбиения исходной совокупности на отдельные группы (классы) однородных объектов, сходных между собой, но имеющих отчетливые отличия друг от друга.

При выборе конкретного метода многомерного анализа данных важное значение имеет наличие или отсутствие обучающей выборки, которая определяет набор эмпирических отношений, устанавливающих статистическую связь $x \to y$, где $x \in \mathbf{X}$, $\mathbf{X}$ – множество объясняющих переменных (predictors), $y \in \mathrm{Y}$, $\mathbf{Y}$ – значения отклика (responses) или "метки" классов. В связи с этим различают две задачи: а) индуктивного распознавания с обучением по прецедентам и б) формирования решающих правил без "учителя".

На рис. 6.1а слева априорная классификация объектов отсутствует, поэтому первая ось главных компонент PC_1 проводится из общих соображений, а именно, через центр тяжести данных и в направлении, совпадающим с наибольшей по длине полуосью эллипсоида рассеяния. Вторая ось PC_2 также проводится через центр распределения перпендикулярно к первой и совпадает по направлению со второй из главных полуосей эллипсоида рассеяния. Эта операция обеспечивает формирование графической метафоры данных с минимально возможными искажениями и в новых ортогональных координатах PC_1 - PC_2, которые оптимизированы относительно распределения исходных данных. В результате становится возможным, например, обозначить разделяющую плоскость между двумя сгущениями объектов.

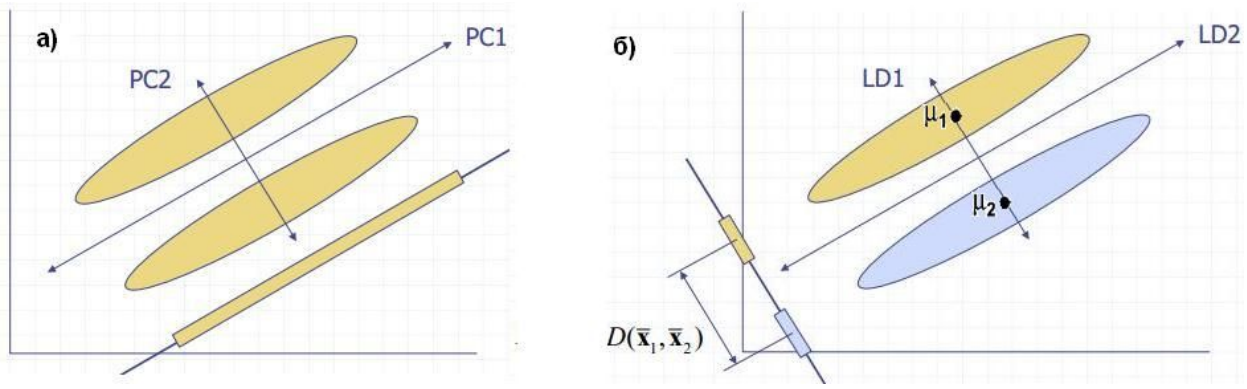


Рис. 6.1. Геометрическая интерпретация методов главных компонент (а) и линейного дискриминантного анализа (б)

На рис. 6.1б справа каждый из объектов эмпирической выборки был отнесен заранее к одному из двух классов. Это дает нам основание провести первую ось дискриминации LD_1 через координаты математических ожиданий μ_1 и μ_2 обоих классов. Вторая ось дискриминации LD_2 проводится перпендикулярно первой и делит расстояние $D(\bar{x}_1, \bar{x}_2)$ между центроидами сгущений в соответствии с принципом "равной удаленности". В многомерном случае через эту ось проходит разделяющая поверхность, уравнение которой может использоваться для распознавания, т.е. прогнозирования предположительных меток классов для объектов экзаменационной последовательности.

Таким образом, при классификации мы ставим задачу выделить некоторые правила, позволяющие отличить между собой совокупности наблюдений, и стремимся оценить, как выделенные классы соотносятся друг с другом, т.е. упорядочить их по взаимному расположению в многомерном пространстве. Здесь прослеживается следующая естественная триада (Кендалл, Стьюарт, 1976, с. 437):

- *кластеризация* или автоматическое разбиение исходного множества объектов на группы по уровню сходства между ними (см. разделы 5.4 – 5.6);
- *статистический анализ* классификационной системы, заданной априори или выделенной в ходе кластеризации: сюда входит изучение отношений эквивалентности между классами, выявление закономерностей и построение совокупности диагностических правил ("дистилляция эталонов") для последующего распознавания;
- *распознавание* (синонимы: прогнозирование или дискриминация) использует классификационные решающие правила для диагностики свойств "новых" объектов, не участвовавших на этапе обработки исходных таблиц.

В дополнение к этим этапам иногда вводят стадию валидации (validation), цель которой – проверка достоверности найденных закономерностей. В дальнейшем мы будем считать валидацию частью второй стадии, поскольку диагностические правила в отрыве от процедур их тщательного тестирования вряд ли достойны отдельного упоминания.

Метод (алгоритм), которым проводят классификацию, называют *классификатором*. Классификатор переводит вектор признаков объекта x в целое число $g = \{1, 2, \ldots, k\}$, соответствующее номеру группы, к которой относится каждый объект. Эта процедура иногда интерпретируется как частный случай общей схемы регрессионного анализа, когда зависимая переменная принимает специальные значения, а в критериях качества модели вместо суммы квадратов остаточных разностей фигурирует функция потерь от неправильной классификации. Хотя к настоящему времени придумано уже несколько десятков конкретных дефиниций функции потерь, наиболее удобной и естественной является $E(r) = F/(T + F)$, т.е. доля ошибок распознавания F при T правильных ответов. При классификации важно понимать, ошибку какого рода важнее минимизировать, поэтому часто вводят систему штрафов, взвешивающих относительную важность отдельных исходов. Например, в юриспруденции, руководствуясь презумпцией невиновности, необходимо минимизировать ошибку 1-го рода – вероятность ложного обвинения. В медицине, при гипотезе "болен", необходимо минимизировать ошибку 2-го рода – вероятность не распознать болезнь.

Стандартной эмпирической методикой тестирования и сравнения алгоритмов классификации, регрессии и прогнозирования является *скользящий контроль* (или кросс-проверка – cross-validation, CV), описанный в разделе 3.4. При этом исходная выборка многократно случайным образом разбивается на q блоков равной длины, после чего каждый блок по очереди становится контрольной выборкой, а объединение всех остальных блоков – обучающей последовательностью. Частным случаем полного скользящего контроля является кросс-проверка с исключением одного объекта (leave-one-out CV), т.е. $q = n$. При этом строится n моделей распознавания по $(n - 1)$ выборочным значениям, а исключенная реализация каждый раз используется для расчета ошибки скользящего контроля, т.е. алгоритм во многом напоминает процедуру "складного ножа" jackknife.

Ошибкой скользящего контроля $CV(r, X^n) = \frac{1}{n}\sum_{i=1}^{n} E(r_k, X_k^{n-k})$ при использовании решающего правила r на исходной выборке X^n называется средняя по всем k разбиениям величина ошибки $E(r)$ на контрольных подвыборках. Если выборка независима, то ошибка скользящего контроля даёт несмещённую и эффективную (Вапник, Червоненкис, 1974) оценку вероятности потерь. Это выгодно отличает её от средней ошибки на обучающей выборке, которая почти всегда оказывается смещённой (оптимистически заниженной), что связано с явлением переобучения. Следует, однако, отметить, что скользящий контроль дает точечную оценку вероятности ошибочной классификации. В настоящее время не существует методов построения на этой основе точных доверительных интервалов для риска, то есть математического ожидания потерь. Иными словами, скользящий контроль эффективно оценивает качество метода распознавания по сравнению с другими аналогичными алгоритмами, но не в состоянии учесть, насколько плоха может оказаться обучающая выборка в конкретном единичном случае.

Разработанные к настоящему времени многочисленные методы машинного распознавания основываются на разнообразных математических платформах поиска наилучшего решения. Статистические методы распознавания используют принцип максимума апостериорной вероятности (оптимальный байесовский классификатор) или алгоритмы восстановления функций плотности распределения вероятности по эмпирической выборке. Линейный или квадратичный *дискриминантный анализ* Фишера основывается на вычислении собственных значений ковариационных матриц и чувствителен к отклонениям от предположений о нормальном законе распределения исходных данных. Метрические методы классификации, такие как алгоритмы *k-ближайших соседей* или потенциальных функций, основаны на вычислении оценок сходства между объектами и опираются на гипотезу компактности сгущений кластеров. Линейные классификаторы, к которым относят *логистическую регрессию* и алгоритм *опорных векторов*, используют принцип максимума правдоподобия или находят оптимальную разделяющую гиперплоскость, обеспечивающую минимальный эмпирический риск ошибки. Некоторые методы классификации, такие как *иерархические деревья решений*, вообще не используют метрик многомерного пространства, а основаны на логических операциях последовательного разделения обучающей выборки на группы на основе установленных зависимостей отклика от независимых переменных

В отличие от классификации, при использовании ординационных методов мы в качестве результата получаем знания, как соотносятся объекты в терминах выбранной метрики, на основании которых стремимся приписать сгущениям точек на ординационной плоскости некоторые классификационные сущности. Большинство методов ординации основано на идеях оптимального целенаправленного проецирования (projecting pursuit) в пространства малой размерности. Такой подход, основанный на минимально возможном искажении исходной взаимной упорядоченности точек, обеспечивает наглядное графическое представление исследуемых объектов на диаграммах с 2 или 3 осями координат. Некоторые методы, такие *самоорганизующиеся карты Кохонена* SOM, являются, в некотором смысле, гибридом ординации и кластеризации.

Наиболее распространенные алгоритмы ординации, такие как анализ *главных компонент* (PCA) или *главных координат* (PCoor), *канонический корреспондентный анализ* (CCA), основаны на операциях с собственными числами и собственными векторами обрабатываемых матриц и предъявляют ряд требований к характеру распределения исходных данных. Весьма перспективным методом считается алгоритм *неметрического многомерного шкалирования* (NMDS, nonmetric multidimensional scaling), который использует различные градиентные методы оптимизации эвристических функционалов качества. Его главным преимуществом является то, что он не требует от исходных данных никаких априорных предположений.

Методы, отмеченные курсивом, будут подробно описаны в последующих разделах.

6.2. Проецирование данных в пространство малой размерности методом PCA

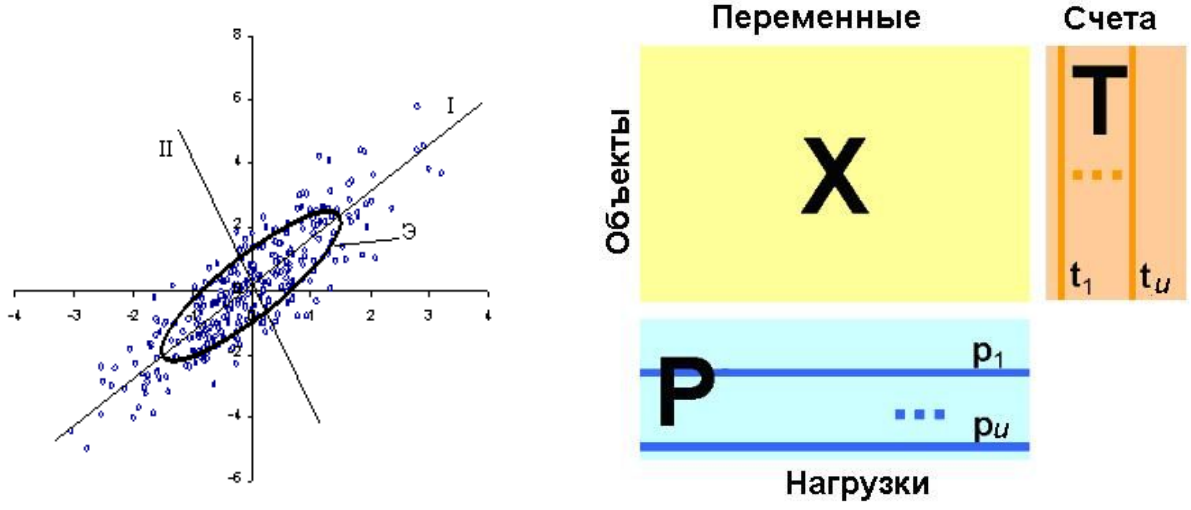
Анализ главных компонент (PCA, principal component analysis) является классическим методом снижения размерности данных, широко используемым в различных областях науки и техники и детально описанным в многочисленных руководствах (Rao, 1964; ter Braak, 1983; Айвазян и др., 1989). В отличие от регрессии, PCA не оценивает одну единственную переменную отклика, а выполняет симметричную обработку всей матрицы наблюдений. При этом алгоритм стремится построить небольшое количество ортогональных плоскостей, ориентируя их относительно максимума вариации точек отображаемых объектов, т. е. при проецировании на эти плоскости вносятся минимально возможные искажения в геометрию исходных данных. Разработанные алгоритмы шкалирования и вращения осей главных компонент позволяют получить удобную для интерпретации факторов ординационную диаграмму.

Пусть имеется m ($m >> 1$) случайных переменных X_1, X_2, …, X_m, имеющих совместное многомерное распределение (не обязательно нормальное), которым соответствует вектор средних $\boldsymbol{\mu}^{m\times 1}$. Ковариационная матрица $\mathbf{S}^{m\times m}$ определяет характер взаимосвязи между этими переменными или их *структурную зависимость*. Метод главных компонент рассматривает в качестве допустимых преобразований всевозможные линейные ортогональные центрированные комбинации переменных X_i:

$$Z_k = \sum_{i=1}^{m} p_{ik}(X_i - \mu_i) \text{ , где } p_{ik} \text{ – пересчетные коэффициенты, } \sum_{i=1}^{m} p_{ik} = 1, \quad k = 1, 2, \ldots, m,$$

из которых выбирается *ортогональная* система векторов, доставляющая максимум критерию информативности (например, доле от суммарной вариабельности исходных признаков – Айвазян и др., 1989). Первой главной компонентой $\mathbf{z}_1$ называется такая линейная комбинация исходных переменных, которая обладает наибольшей дисперсией (рис. 6.2а). В свою очередь, каждая следующая k-я главная компонента $\mathbf{z}_k$ ($k = 2, \ldots, m$) не коррелированна с k - 1 предыдущими главными компонентами и имеет наибольшую дисперсию по сравнению с остальными. Ранжирование по дисперсии осей найденных латентных переменных позволяет выполнить поиск такой u-мерной системы координат ($m > u$), которая содержит сжатое описание структурной зависимости исследуемой системы признаков $\mathbf{X}$, определенной в $\mathbf{S}^{m\times m}$, небольшим числом u факторов и без существенной потери информации.

Результатом PCA-анализа конкретной матрицы наблюдений $\mathbf{X}$ размерностью ($n\times m$), где n – число наблюдаемых объектов, m – число независимых переменных, является матрица $\mathbf{T}$ *счетов* (scores) размерностью $n\times u$, содержащая проекции исходных точек выборки $\mathbf{X}$ в новом u-мерном базисе (см. рис. 6.2б).



а) б)

Рис. 6.2. Слева (а): пример двумерного распределения точек; I, II – оси главных компонент; Э – эллипсоид рассеяния. Справа (б): разложение матрицы наблюдений по главным компонентам

Другая матрица **P** размерностью ($u \times m$) содержит *нагрузки* (loadings), обеспечивающие пересчет данных из m-мерного пространства исходных переменных в u-мерное пространство главных компонент. Каждая k-я строка матрицы **P** состоит из оценок коэффициентов p_{ik}, показывающих долю участия каждой i-й переменной из $\mathbf{x}_1$, …$\mathbf{x}_m$ в формировании k-й главной компоненты. Фактически это – проекция признака $\mathbf{x}_j$ на новую k-ю ось. Анализируя таблицу нагрузок, можно понять, какие исходные переменные связаны между собой, а какие независимы, и объяснить предметный смысл k-го фактора.

В контексте этих матриц термин "главная компонента" может относиться как к счетам (т.е. совокупности выборочных проекций на ось PC_i), так и к нагрузкам (т.е. ранжированному набору вкладов исходных переменных, связанных с осью PC_i).

Произведение матриц **T** и **P** является некоторой аппроксимацией $\widetilde{\mathbf{X}}$ анализируемых данных **X** в редуцированном u-мерном пространстве: $\widetilde{\mathbf{X}} = \mathbf{T}_u \mathbf{P}_u^t$. При $u = m$ приближение становится точным и $\mathbf{X} = \widetilde{\mathbf{X}}$. Потеря информации от снижения размерности описывается матрицей остатков $\mathbf{E} = \mathbf{X} - \widetilde{\mathbf{X}}$, компоненты которой могут быть использованы для оценки остаточных дисперсий (погрешностей) любого i–го объекта и доли объясненной дисперсии R:

$$d_i = \frac{1}{n}\sum_{j=1}^{m} e_{ij}^2 ; \qquad R = 1 - \sum_{i=1}^{n} d_i \Big/ \sum_{i=1}^{n}\sum_{j=1}^{m} x_{ij}^2 .$$

Снижение размерности исходного пространства методом PCA можно представить как последовательный, итеративный процесс, который можно оборвать на любом шаге u. Вследствие *ортогональности* системы координат главных компонент (т.е. фактически, их взаимной независимости) нет необходимости перестраивать матрицы счетов **T** и нагрузок **P** при изменении числа компонент: к ним просто прибавляются или отбрасываются очередные столбцы (строки). При этом важно знать, когда следует остановиться, т.е. определить "правильное" число главных компонент u. Если это число слишком мало, то описание данных будет не полным. С другой стороны, избыточное число главных компонент приводит к переоценке, т.е. к ситуации, когда моделируется шум, а не содержательная информация.

Главные компоненты определены таким образом, что достигается оптимум трех эквивалентных критерия (Wehrens, 2011):

- каждый столбец матрицы счетов имеет дисперсию, максимально возможную из всех линейных комбинаций;
- сумма евклидовых расстояний между проекциями точек в редуцированном пространстве максимальна;
- реконструкция **X** насколько возможно близка к оригиналу и $||\mathbf{X} - \widetilde{\mathbf{X}}||$ минимально.

Существуют различные методы оценки матрицы коэффициентов **P**, удовлетворяющих этим критериям, такие, как сингуля́рное разложе́ние ковариационной матрицы (singular value decomposition, SVD), EM-алгоритм, ищущий оценку максимального правдоподобия.

Наиболее распространенный метод преобразования $\mathbf{XP} \rightarrow \mathbf{T}$ основан на вычислении последовательности собственных значений $\lambda_1 > \lambda_2 > \ \ldots > \lambda_m$ квадратной симметричной матрицы $\mathbf{S} = \mathbf{X}^t\mathbf{X}$ вторых или смешанных моментов. Если столбцы матрицы **X** центрированы ($\sum_{i=1}^{n} x_{ij} / n = 0$), то матрица **S** является ковариационной, а если еще и выполнена нормировка ($\sum_{i=1}^{n} x_{ij}^2 / n = 1$), то **S** становится корреляционной матрицей.

Максимальная дисперсия проекций на 1-ю главную компоненту достигается для собственного вектора $\mathbf{p}_1$ выборочной матрицы **S**, которому соответствует максимальное собственное значение λ_1. Рассматривая аналогично проекции на новые направления, находится наилучшее m-мерное линейное подпространство, которое определяется набором $\mathbf{P}^{m \times m}$ собственных векторов матрицы **S**, отвечающих m найденным собственным значениям $\boldsymbol{\lambda}$. Ортогональность полученной системы координат достигается соблюдением условия $\mathbf{P}^t\mathbf{P} = \mathbf{I}$, где **I** – единичная матрица.

В общем виде для полученных матриц имеет место следующее соотношение:

$$\mathbf{T}^t\mathbf{T} = \mathbf{P}'\mathbf{S}\mathbf{P} = \mathbf{\Lambda}, \quad \text{где } \mathbf{\Lambda} = \text{diag}\{\lambda_1,\ldots,\lambda_m\},$$

т.е. дисперсии столбцов матрицы **T** соответствуют собственным значениям λ_j матрицы **S,** являющимся диагональными элементами матрицы **Λ**. В результате выполненных расчетов матрица **T** имеет те же размеры, что и **X**, однако ее столбцы не коррелируют между собой.

Обычно исследователь ориентируется на последовательность собственных значений λ_1, ..., λ_m и решает, сколькими главными осями ему стоит ограничиться. Это решение может быть полностью произвольным: например, для интерпретации считается достаточным число осей *u*, объясняющих 75 % дисперсии в исходных данных, т.е. $R^2 = \sum_l^u \lambda_l / \sum_l^m \lambda_l = 0.75$, где R^2 – некоторый аналог коэффициента детерминации в регрессионном анализе. Критерий Кайзера-Гуттмана рекомендует оставить только те главные компоненты, собственные значения которых превышают среднее $\sum_k^u \lambda_k / u$. Более осмысленный подход основан на модели "разбитой палочки" (broken stick model - MacArthur, 1957)[11], которая описывает процесс случайного разделения ресурса *L* на *A* частей. Если доля дисперсии, объясненной λ_k, не превышает доли, рассчитанной по модели broken stick, то *k*-я главная компонента (и все последующие) считаются тривиальными, поскольку их появление носит случайный характер.

В качестве примера выполним анализ методом главных компонент массива геоботанических описаний, полученных со 159 пробных площадок в дельте р. Волга (пример [П2], см. также разделы 4.5, 5.3 и 5.5). В большинстве руководств по статистической обработке предлагается предварительно стандартизовать и центрировать исходные данные с использованием *z*-преобразования: $z_i = (x_i - \bar{x})/s^2$, где $\bar{x}$ – среднее и s – стандартное отклонение. После такой трансформации все переменные, имеющие в нашем случае значения биомасс 22 видов травянистых растений, будут иметь однородный характер с нулевыми средними и единичными дисперсиями, т.е. каждый вид приобретает равноправный статус, вне зависимости от того, является ли он экологическим доминантом или встречается редкими единичными экземплярами. Однако, если рассчитать последовательность собственных значений ковариационной матрицы **S** (22×22), то мы обнаружим, что объясняемая дисперсия равномерно разложилась на большое число главных компонент, каждая из которых связана через нагрузки только с 1-2 видами:

Собственные	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8
значения:	1.515	1.325	1.286	1.182	1.110	1.100	1.060	1.041
Отношение R^2:	0.105	0.185	0.261	0.325	0.381	0.436	0.488	0.537

Если же не выполнять предварительного *z*-преобразования матрицы **X**, то мы оказываемся в затруднении противоположного свойства: все статистически значимые нагрузки p_i связаны в основном с первой главной компонентой:

Собственные	λ_1	λ_2	λ_3	λ_4	λ_5	λ_6	λ_7	λ_8
значения:	62.6	25.30	20.18	18.44	16.07	15.08	13.25	12.35
% дисперсии:	0.597	0.097	0.062	0.051	0.039	0.034	0.026	0.023
Отношение R^2:	0.597	0.695	0.757	0.809	0.848	0.883	0.909	0.933

На рис. 6.3 легко увидеть, что только для λ_1 доля объясненной дисперсии превышает случайное значение, рассчитанное по модели "разбиваемой палочки", тогда как критерий Кайзера-Гуттмана, не привычный к столь экстремальным ситуациям, оптимистично советует увеличить число осей главных компонент вплоть до 4.

[11] Проф. В.Н. Максимов обратил внимание, что устоявшийся перевод *broken stick* как «разломанный стержень» не вполне точен, поскольку не отражает случайный характер процесса. Более содержательна метафора «стеклянной палочки, которая разбивается, упав на пол».

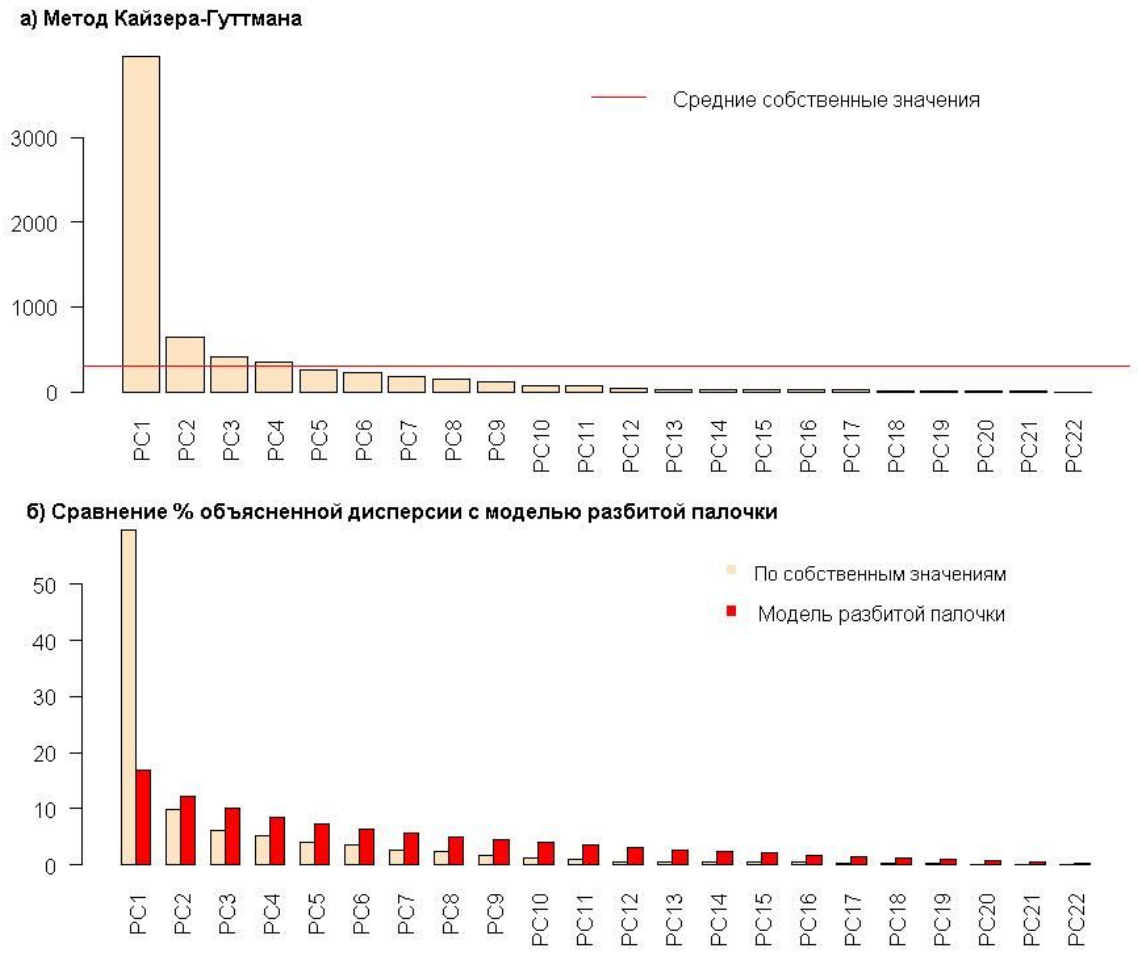


Рис. 6.3. Оценка числа нетривиальых главных компонент различными методами

Непосредственно для целей ординации оценка числа информационно значимых главных компонент имеет вспомогательное значение, поскольку ключевым подходом является проецирование точек данных на плоскость, координатные оси которой PC_1–PC_2 рассчитываются как линейные комбинации факторных нагрузок, основанных на первых двух собственных значениях и их собственных векторах. На совместной ординационной диаграмме (или биплоте – см. рис. 6.4) показываются одновременно точками местоположение объектов в новой системе координат на основе матрицы счетов **T** и стрелками – проекции каждой исходной переменной $\mathbf{x}_j$ на основе матрицы нагрузок **P**.

Для удобства визуализации точек на ординационной диаграмме выполняется шкалирование, определяющее способ проецирования и масштаб соотношений между графическими объектами. Поскольку нет единственного пути построения оптимального изображения, шкалирование проводят различными методами (Legendre, Legendre, 1998):

° *шкалирование 1* или дистанционный биплот, который приведен к масштабу "единичной длины" собственных значений; расстояния между объектами приближены к евклидовым дистанциям в многомерном пространстве, а углы между векторами переменных не имеют статистического смысла;

° *шкалирование 2* или корреляционный биплот, приведенный к масштабу $(\lambda_k)^{0.5}$, в котором расстояния между объектами не имеют смысла евклидовых дистанций, но углы между векторами переменных отражают меру корреляции между ними.

Чтобы уточнить, какая экологическая реальность скрыта в найденных главных компонентах, проводится анализ матрицы факторных нагрузок **P**, элементы которой тем больше (по абсолютной величине), чем больше статистическая связь исходных переменных с каждой из новых осей ординации – см. табл. 6.1.

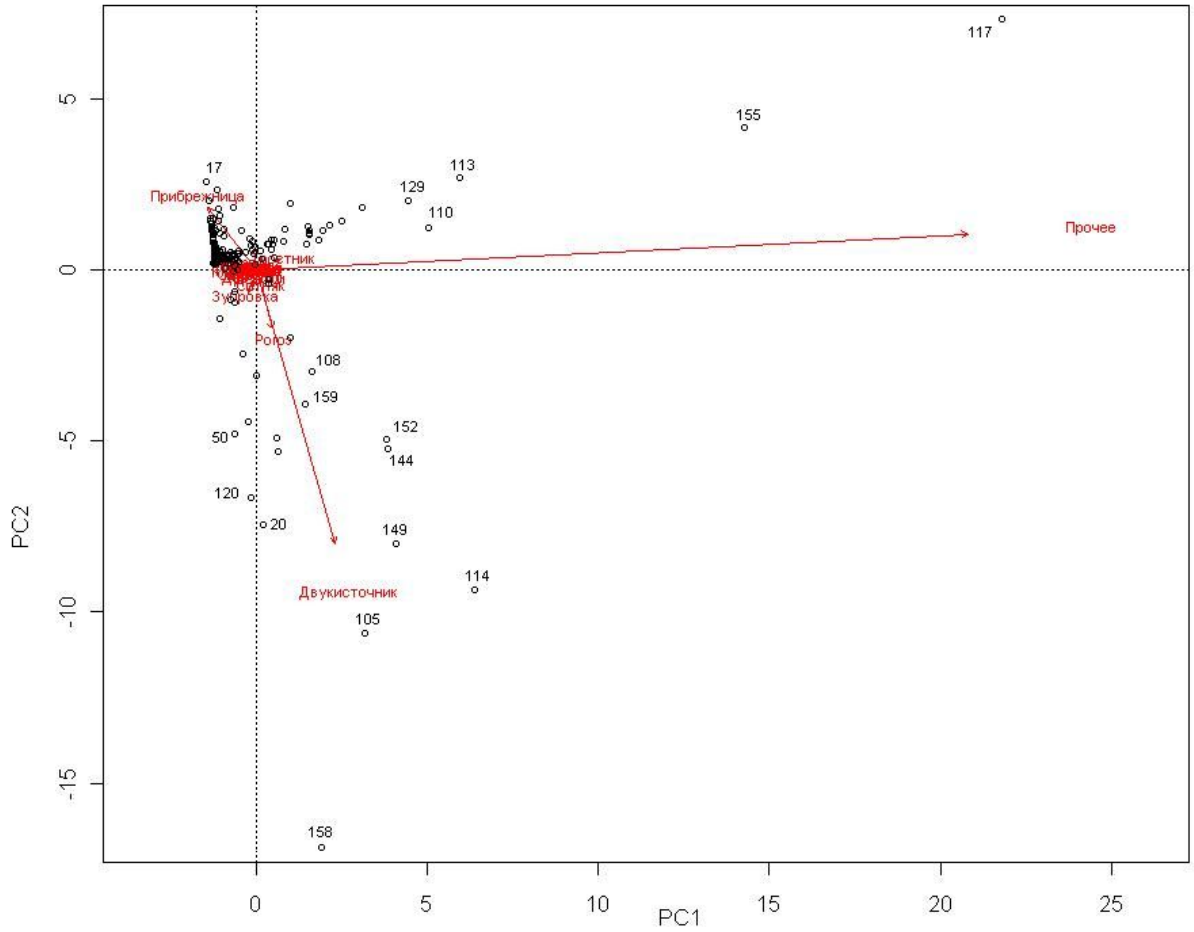


Рис. 6.4. Ординационный биплот геоботанических описаний (метод шкалирования 2)

Таблица 6.1. Значения нагрузок (*P*) и статистической значимости корреляций (*p*) различных видов травянистой растительности с первыми 4-мя главными компонентами

Оси главных компонент	PC1		PC2		PC3		PC4	
	P	*p*	*P*	*p*	*P*	*p*	*P*	*p*
Рогоз	**0.557**	**0.017**	-2.007	0.185	0.779	0.352	-6.985	0.389
Петросимония супр.	**-0.136**	**0**	0.070	0.233	-0.039	0.476	-0.037	0.442
Петросимония	**-0.405**	**0**	0.213	0.244	-0.219	0.415	-0.149	0.444
Скрытница	-0.351	0.105	-0.148	0.428	-2.070	0.192	0.188	0.452
Солерос	**-0.115**	**0.036**	0.037	0.331	-0.150	0.22	-0.025	0.465
Тростник	0.862	0.106	0.343	0.305	-0.350	0.261	0.092	0.458
Двукисточник	**2.734**	**0**	-9.397	0.167	1.585	0.271	1.629	0.383
Солодка голая	-0.051	0.213	0.111	0.159	0.098	0.294	-0.008	0.476
Прибрежница	**-1.645**	**0**	2.150	0.163	6.702	0.23	0.868	0.458
Клубнекамыш	**-0.477**	**0.04**	-0.082	0.479	-2.664	0.177	0.601	0.419
Девясил	-0.057	0.319	-0.238	0.22	-0.479	0.228	0.161	0.429
Ситняк болотный	0.191	0.083	-0.408	0.199	-0.205	0.341	0.219	0.4
Дербенник лозный	0.009	0.255	0.004	0.471	-0.045	0.242	0.007	0.409
Алтей	**-0.100**	**0.035**	-0.045	0.432	-0.200	0.257	0.046	0.422
Марна	-0.004	0.452	-0.009	0.474	-0.278	0.216	0.039	0.439
Лебеда копьелист.	-0.072	0.067	-0.020	0.473	-0.356	0.186	0.046	0.432
Череда	0.000	0.523	-0.001	0.473	0.000	0.575	0.000	0.608
Пырей	**-0.256**	**0**	0.165	0.289	-0.532	0.203	-0.036	0.504
Дербенник иволист.	0.122	0.133	-0.255	0.413	0.051	0.481	0.102	0.498
Шведка	**-0.224**	**0**	0.119	0.206	0.048	0.382	-0.033	0.453
Зубровка	-0.258	0.101	-0.794	0.226	-1.723	0.225	0.491	0.459
Прочие	**24.470**	**0**	1.227	0.143	0.157	0.38	0.047	0.476

Анализ PCA, представленный на рис. 6.4 и табл. 6.1 показывает, что представленные геоботанические описания являются прекрасным примером данных с существенно "размытой" структурой. Напомним, что PCA-анализ этого примера мы проводили без центрирования матрицы **X**, поэтому первая главная компонента включала как основную долю не слишком явно выраженных корреляций между исходными переменными, так и смещение среднего многомерного распределения анализируемой выборки (в терминах факторного анализа эта ситуация известна как "общий фактор"). Как показано на биплоте рис. 6.4, облако данных в целом представляет небольшой сгусток сигарообразной формы. Причем статистическая значимость второй и последующих ортогональных осей уже не прослеживается, а наибольшая изменчивость переменных связана с "Прочими видами", которыми и определяется специфика фитоценозов.

Целостный подход к оценке статистической значимости связи исходных переменных с главными компонентами может быть реализован с использованием рандомизации и бутстрепа. Здесь возможны два основных алгоритма. Классический непараметрический бутстреп будет реализован, если по схеме "выбора с возвращением" выбирается случайная совокупность из n порядковых номеров строк исходной матрицы наблюдений **X** и из этих строк формируется бутстрепированная таблица $\mathbf{X}_b^*$. Поскольку порядок следования переменных в строках не изменяется, а, следовательно, связи между переменными не нарушены, матрицу $\mathbf{X}_b^*$ назовем *скоррелированной перевыборкой*. При большом числе итераций бутстрепа ($B \to \infty$) оценки ковариации (корреляции) s_{ij}^* каждой пары переменных будут изменяться от минимального до максимально возможного значения, которые определяются внутренней структурой исходной выборки. Тем самым мы оцениваем потенциальную информационную важность k-й компоненты в диапазоне самых оптимистичных и самых пессимистичных комбинаций строк данных.

Другой алгоритм, выполняющий рандомизацию, случайно перемешивает еще и сами элементы в пределах каждого столбца матрицы **X** и тем самым нарушает естественные связи между парами переменных, т.е. мы получаем *нескоррелированную перевыборку* $\mathbf{X}_r^*$. Если также использовать случайный выбор с возвращением, то вариация собственных значений λ_k ковариационных матриц $\mathbf{S}_r^*$, рассчитанных по перевыборкам $\mathbf{X}_r^*$, оценивает изменчивость структуры при справедливости нулевой гипотезы, т.е. для нуль-модели данных, где взаимосвязь всех исходных факторов носит случайный характер.

Для оценки статистической значимости вклада j-й переменной в k-ю главную компоненту на каждой итерации рандомизации вычисляется матрицу нагрузок $\mathbf{P}_r^*$ и подсчитывается число случаев b, когда значение нагрузки P_{jk}^*, найденное для матрицы $\mathbf{X}_r^*$, превысило бы аналогичное значение для эмпирической матрицы **X**. Величины $p = b/B$ для рассматриваемого примера представлены в табл. 6.1.

В. Пиллар (Pillar, 1999) предложил уточненную схему тестирования метрических ординаций (CA, PCA или PCoA) с использованием бутстреп-метода:

- генерируется большое число B (например, B = 1000) псевдовыборок на основе случайных комбинаций из столбцов исходной матрицы **X**;
- полученные бутстреп-матрицы подвергаются ординации по одинаковой схеме с последующей "прокрустовой стандартизацией" (Procrustean adjustment), т. е. путем вращения и корректировки нагрузок структуры приводятся к сопоставимой форме;
- для каждой псевдовыборки рассчитывается средний коэффициент корреляции θ^*_1 факторных весов 1-й главной ординационной оси со значениями натуральных признаков;
- значения показателей, представленных в строках псевдоматрицы, многократно случайным образом перемешиваются, рассчитывается средний коэффициент корреляции θ^o_1 для рандомизированной структуры и проверяется выполнение условия $\theta^o_1 > \theta^*_1$;
- вычисляется вероятность ошибки 1-го рода $p(\theta^o_1 > \theta^*_1) = (b + 1) / (B + 1)$, где b – число случаев, когда коэффициент корреляции для рандомизированных данных оказался больше, чем для эмпирических;
- аналогичные вычисления проводятся для остальных осей ординации.

Вероятность $p(\theta^{o}_{1} > \theta^{*}_{1})$, рассчитанная бутстреп-методом по схеме Пиллара, является индикатором устойчивости ординационной структуры изучаемой экосистемы по сравнению с ее нуль-моделью. С использованием программы MULTIV нами (Шитиков и др., 1912) было проанализировано на одном и том же исходном материале, как изменяется надежность результатов ординации в зависимости от таких ключевых параметров расчета как размерность признакового пространства, тип обрабатываемых данных и формулы для расчета меры сходства. К сожалению, нам не удалось найти функции, реализующей алгоритм Пилара в статистической среде R.

Рассмотрим другой пример [П2], который мы использовали в разделах 5.4 и 5.6 при кластеризации 13 участков рек Сок-Байтуган по 129 признакам, определяющим отдельные таксоны бентофауны водотоков. В этом случае R-способ, основанный на вычислении ковариационной матрицы **S** между переменными, технически не может быть использован для оценки собственных значений. Если количество наблюдений n меньше или равно числу переменных m, то матрица **S** порядка m имеет только $(n - 1)$ независимых строк (столбцов) и мы получим $[m - (n - 1)]$ неопределенных собственных значений. Однако Рао (Rao, 1964) была разработана техника транспонирования исходной матрицы **X** и вычисления матрицы корреляции между объектами (то есть Q-способ), что позволило реализовать в этих условиях PCA-анализ и получить статистически корректную интерпретацию итогов ординации в редуцированном пространстве $u = \min(n - 1, m)$.

С использованием Q-анализа мы можем вычислить 12 главных компонент, что вполне достаточно для ординации как участков, так и таксонов макрозообентоса. Выполним предварительно оценку доверительных интервалов собственных значений λ_k с использованием непараметрического бутстрепа. Можно отметить (см. табл. 6.2), что незначительные изменения, вносимые бутстрепом в исходную таблицу **X**, весьма серьезно сказываются на вариации вычисляемых собственных значений и перераспределении относительной роли главных компонент. Интересным феноменом, требующим специального осмысления, является несимметричность доверительных интервалов, найденных методом процентилей, причем для накопленной доли объясненной дисперсии они даже не всегда "накрывают" эмпирическое значение R^2.

Таблица 6.2. Оценка верхних (CI_{low}) и нижних (CI_{high}) границ 95% доверительных интервалов собственных значений λ_k и долей объясненной вариации первыми 4-мя главными компонентами по данным гидробиологической съемки на реках Сок-Байтуган; m_{obs} – показатели, рассчитанные по эмпирическим данным

	Собственные значения λ_k			Объясненная дисперсия, %			Накопленная доля R^2, %		
	m_{obs}	CI_{low}	CI_{high}	m_{obs}	CI_{low}	CI_{high}	m_{obs}	CI_{low}	CI_{high}
λ_1	73	42	120	41	32	61	41	32	61
λ_2	22	20	41	12	11	26	53	53	76
λ_3	16	14	27	9	8	17	62	65	87
λ_4	13	11	18	7	6	12	69	74	94

У.Ревелле (Revelle, Rocklin, 1979) обосновал концепцию Очень Простой Структуры (Very Simple Structure) и предложил критерий VSS, оценивающий насколько сильно анализируемые данные отклоняются от нее. На практике, осуществляя снижение размерности, критерий VSS используется не только для того, чтобы выбрать наилучший метод вращения факторов, но и для оценки числа информационно значимых главных компонент, дальнейшее увеличение которых неоправданно с точки зрения конкретной конфигурации данных. Эта идея легко обобщается с использованием бутстреп-процедуры, предложенной С.В. Петровым на тематическом сайте, посвященном статистической среде R (http://p2004r.blogspot.ru/2011_04_01_archive.html).

Несколько модифицируем этот скрипт, воспользовавшись тем, что гидробиологические данные имеют счетный характер (число проб, в которых встретился каждый вид организмов). В качестве альтернативной нуль-модели VSS, в которой внутренние корреляционные связи разрушены, используем процедуру **r2dtable** (Patefield,

1989), которая требует, чтобы маргинальные суммы по строкам и столбцам бутстрепированной матрицы $\mathbf{X}_r^*$ в точности соответствовали бы аналогичным суммам эмпирической матрицы **X**. Сами значения численностей в ячейках могут принимать случайные целочисленные значения. Далее построим совместный график изменения собственных значений, полученных по исходной матрице наблюдений и на основе каждой ее перевыборки $\mathbf{X}_r^*$, а искомый оптимум числа главных компонент найдем, когда эмпирическая кривая сольется с пучком линий случайных реализаций – см. рис. 6.5.

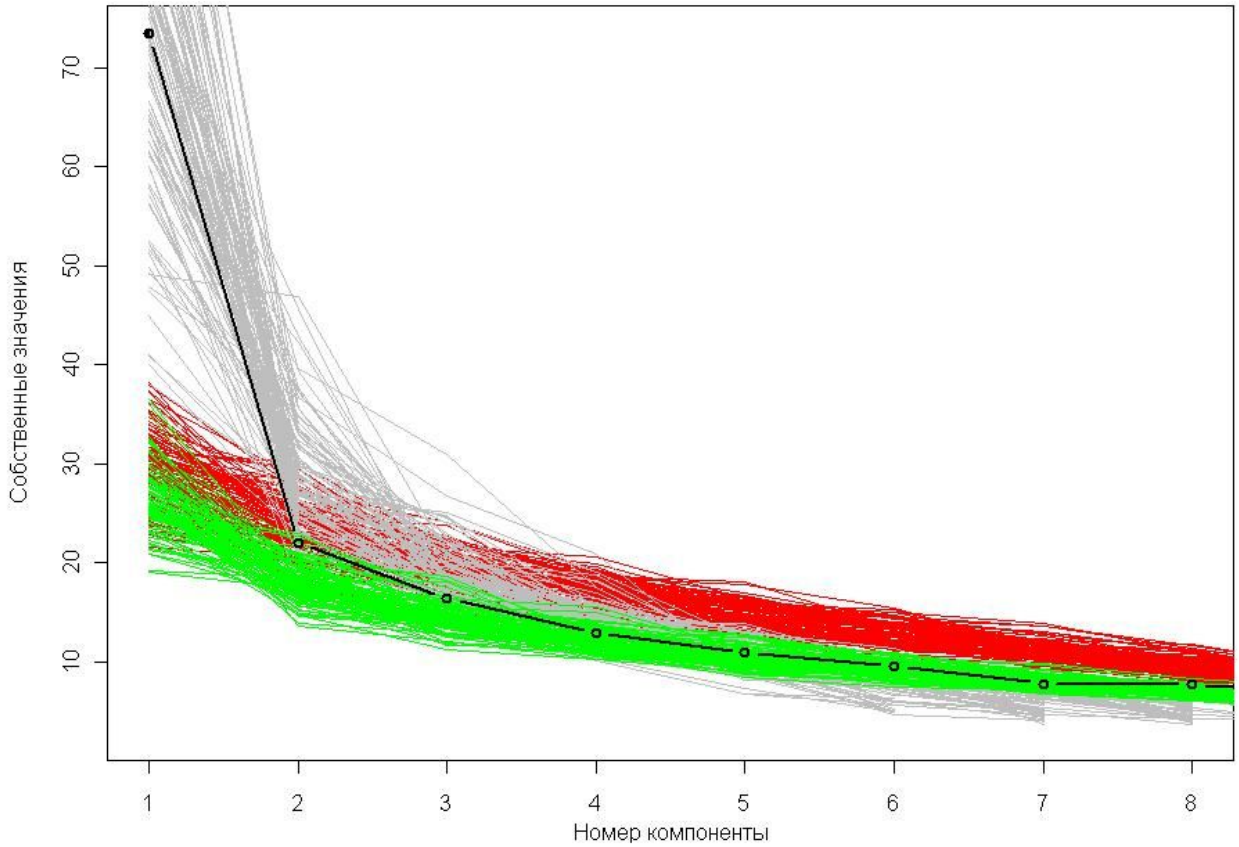


Рис. 6.5. Изменение собственных значений для эмпирической матрицы (линия черного цвета), коррелированных (пучок линий серого цвета) и нескоррелированных (пучок линий красного цвета) перевыборок бутстрепа, а также полученных на основе нуль-модели r2dtable (пучок линий зеленого цвета); оценка числа информативных главных компонент – между 2 и 3

Возможно, у читателя появится мысль, что, если основной целью исследования является построение ординационных биплотов с использованием только первых двух главных компонент, то оценка информационной значимости остальных ортогональных осей не имеет практической ценности. Однако во всех случаях полезно знать, какая доля необъясненной изменчивости данных осталась за бортом нашего графика.

Особенности структуры сообществ донных организмов в реках Сок-Байтуган хорошо прослеживаются на двухмерных ординационных диаграммах: главная ось PC1 ординации объектов (рис. 6.6) позволяет легко выделить группы участков водотока, соответствующих устьевой зоне (С-12, С-13), среднему течению (С-4 – С-9) и верховьям (остальные участки), а вторая ось PC2 показывает изменение видового состава бентоса под влиянием специфических местных условий.

Аналогичная ординационная диаграмма переменных (рис. 6.7) позволяет выделить ассоциации таксонов донных организмов, характерных для каждого местообитания, что в конечном итоге определяется совокупностью факторов окружающей среды. Функции ordipointlabel() и ordilabel() пакета vegan дают возможность расположить на биплоте метки данных наилучшим образом и задать последовательность их визуализации: в верхнем слое диаграммы показаны экологически важные виды с наибольшей частотой встречаемости.

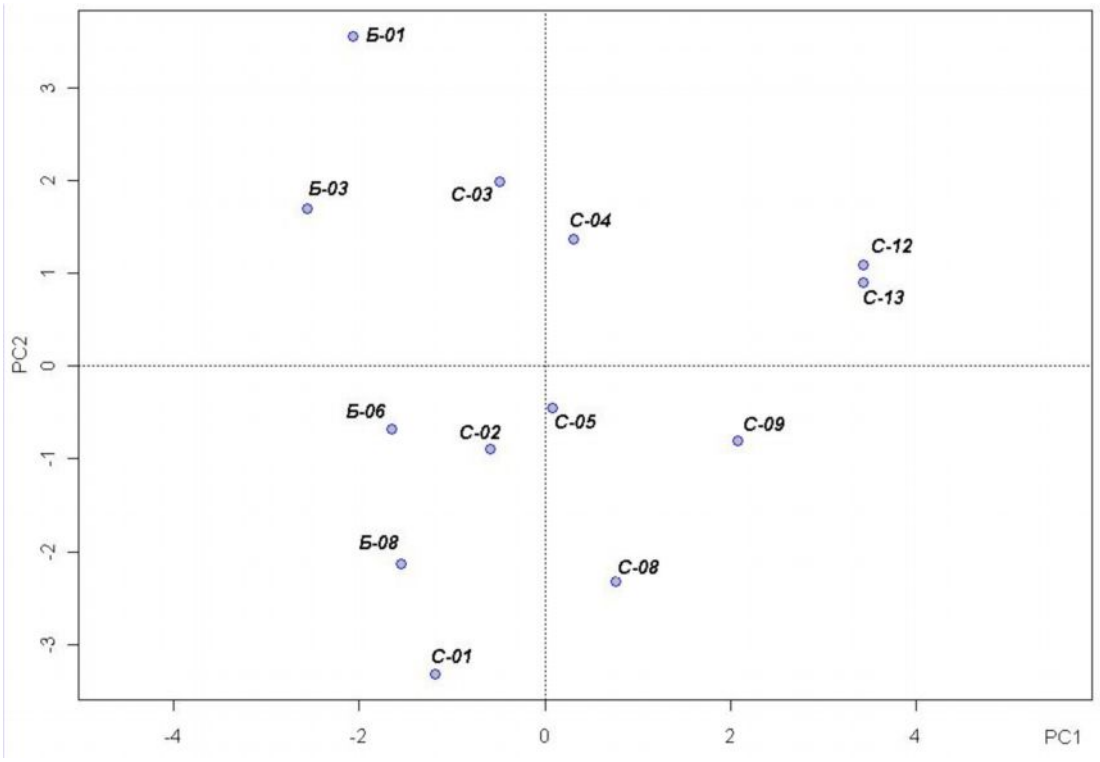


Рис. 6.6. Ординация участков рек Сок-Байтуган по видовому составу донных сообществ

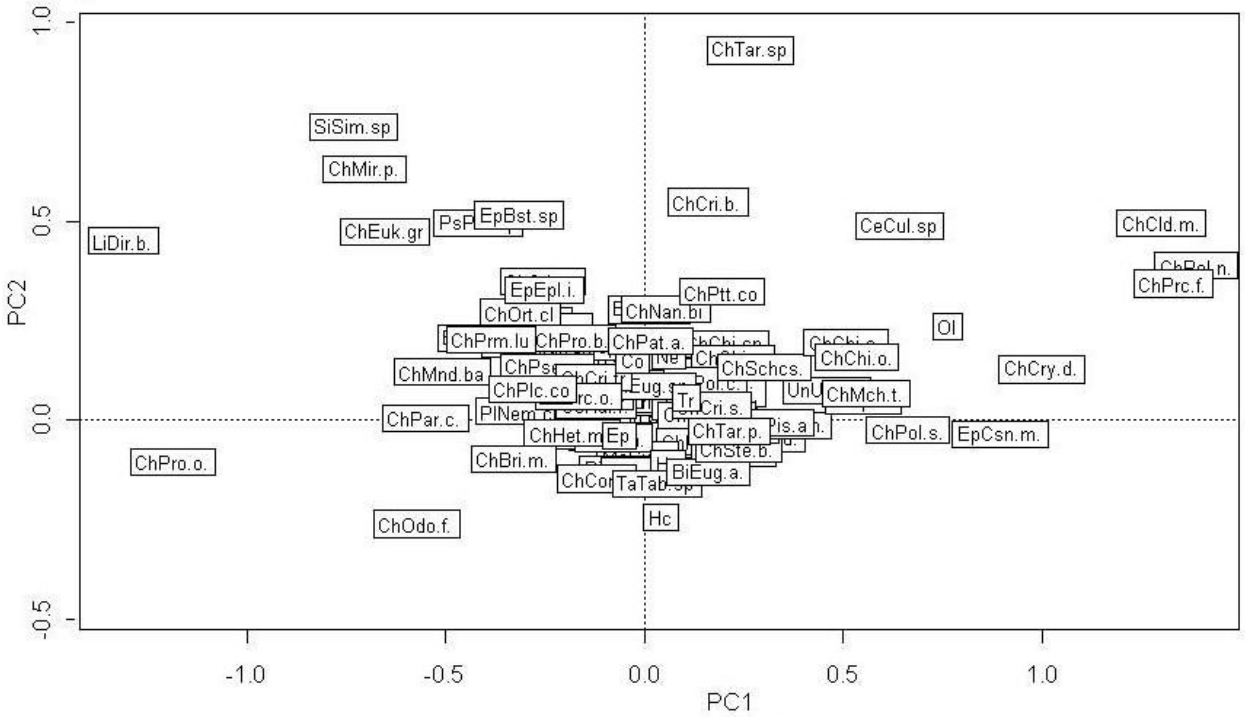


Рис. 6.7. Ординация видового состава донных сообществ рек Байтуган-Сок

К разделу 6.2:

```
#  I. Анализ главных компонент  геоботанических описаний
# Загрузка таблицы видов из предварительно подготовленного двоичного файла
library(vegan)  ; load(file="Fito_Full.RData");  Species <- Species[,-8]
# Для стандартизации переменных используем функцию decostand и z-scores (m=0, sd=1)
Species_stand<-decostand(Species,"standardize")
# 1. Использование функции princomp
# Вариант со стандартизованными переменными
prin_comp<-princomp(Species_stand) ; summary(prin_comp) ; ordiplot(prin_comp, type="text")
```

```
# Вариант с натуральными переменными
prin_comp<-princomp(Species) ; summary(prin_comp) ; ordiplot(prin_comp, type="text")
# 2. Использование функции function prcomp (получаем те же значения)
prin_comp<-prcomp(Species,retx=FALSE,center=FALSE) ; summary(prin_comp)
ordiplot(prin_comp, type="text")
# 3. Использование функции rda (Аргумент scale=TRUE вызывает стандартизацию переменных)
Spec.pca <- rda(Species, scale=FALSE) ; summary(Spec.pca, scaling=1)
summary(Spec.pca, scaling=2)
par(mfrow=c(1,2))  # Два PCA биплота: тип шкалирования 1 и 2
source("cleanplot.pca.R")  # Используем из этого файла функцию biplot.rda()
biplot.rda(Spec.pca, scaling=1, main="PCA - шкалирование 1", type = "text")
biplot.rda(Spec.pca, scaling=2, main="PCA - шкалирование 2", type = "text")
# Оценка статистической значимости связи исходных переменных с главными компонентами
sigpca2<-function (x, permutations=1000, ...) {  # Функция, выполняющая рандомизационный тест
   pcnull <- princomp(x, ...) ;  res <- pcnull$loadings
   out <- matrix(0, nrow=nrow(res), ncol=ncol(res)) ;   N <- nrow(x)
   for (i in 1:permutations) {
       pc <- princomp(x[sample(N, replace=TRUE), ], ...)
       pred <- predict(pc, newdata = x) ;  r <-  cor(pcnull$scores, pred)
       k <- apply(abs(r), 2, which.max)
       reve <- sign(diag(r[k,])) ; sol <- pc$loadings[ ,k] ; sol <- sweep(sol, 2, reve, "*")
       out <- out + ifelse(res > 0, sol <=  0, sol >= 0) }
   out/permutations }
sigpca2(Species, permutations=1000)
# Оценка необходимого числа главных компонент
ev <- Spec.pca$CA$eig ; ev[ev > mean(ev)]  # Критерий Кайзера-Гуттмана
# Модель разбиваемой стеклянной палочки
n <- length(ev) ; bsm <- data.frame(j=seq(1:n), p=0); bsm$p[1] <- 1/n
for (i in 2:n) {bsm$p[i] = bsm$p[i-1] + (1/(n + 1 - i))} ; bsm$p <- 100*bsm$p/n
# Столбиковая диаграмма с данными по этим двум методам
par(mfrow=c(2,1)) ; barplot(ev, main="Собственные значения", col="bisque", las=2)
abline(h=mean(ev), col="red")
legend("topright", "Средние собственные значения", lwd=1, col=2, bty="n")
barplot(t(cbind(100*ev/sum(ev),bsm$p[n:1])), beside=TRUE, main="% дисперсии",
                                  col=c("bisque",2), las=2)
legend("topright", c("% собственных значений", "Модель разбитой палочки"),pch=15,
                                   col=c("bisque",2), bty="n")
#  II.  Анализ главных компонент  гидробиологических данных
library(xlsReadWrite)  #  Загрузка данных из файла Excel
Bentos <- read.xls("Сок1.xls", sheet = 1, rowNames=TRUE) ; Bentos[is.na(Bentos)] <- 0
Bentos <- t(Bentos) ; Spec.pca <- rda(Bentos, scale=FALSE) ; summary(Spec.pca, scaling=1)
#  Оценка доверительных интервалов собственных значений бутстреп-методом
N.boot <- 1000 ; nObs = nrow(Bentos) ; nEig = 4
ci.level = 0.95 ; prob.low = (1-ci.level)/2 ; prob.high = 1-prob.low
boot.eigenval <- boot.eigenprop <- boot.eigencum <-  matrix(0, nrow=N.boot, ncol=nEig)
for (iter in 1:N.boot)  { i = sample(1:nObs, replace=TRUE)
    boot.pca <- rda(Bentos[i,], scale=FALSE) ; boot.eigenval[iter,] = boot.pca$CA$eig[1:nEig]
    boot.eigenprop[iter,] = boot.pca$CA$eig[1:nEig]/sum(boot.pca$CA$eig)
    boot.eigencum[iter,] = cumsum(boot.pca$CA$eig[1:nEig]/sum(boot.pca$CA$eig)) }
CumSum <- cumsum(Spec.pca$CA$eig[1:nEig]/sum(Spec.pca$CA$eig)) ;
Eval <-  matrix(0, nrow=nEig, ncol=9)
colnames(Eval) <- c("Собств. знач.", "Нижн.ДИ","Верхн.ДИ","Объясн.дисп", "Нижн.ДИ",
                    "Верхн.ДИ", "Накоп.доля", "Нижн.ДИ","Верхн.ДИ")
for (k in 1:nEig) {
   Eval[k,1]=Spec.pca$CA$eig[k]
   Eval[k,2] = quantile(boot.eigenval[,k], probs=prob.low, na.rm = TRUE)
   Eval[k,3] = quantile(boot.eigenval[,k], probs=prob.high)
   Eval[k,4]=Spec.pca$CA$eig[k]/sum(Spec.pca$CA$eig)
   Eval[k,5] = quantile(boot.eigenprop[,k], probs=prob.low)
   Eval[k,6] = quantile(boot.eigenprop[,k], probs=prob.high, , na.rm = TRUE)
   Eval[k,7]=CumSum[k]; Eval[k,8]=quantile(boot.eigencum[,k],probs=prob.low,,na.rm=TRUE)
   Eval[k,9] = quantile(boot.eigencum[,k], probs=prob.high, , na.rm = TRUE) } ; Eval
#  Функция создания скоррелированной перевыборки  (бутстреп)
boot  <- function(data) data[sample(1:nrow(data),size=nrow(data),replace=T),]
```

```
# Функция создания перевыборки с «разрушенной» корреляцией (рандомизация)
boot2 <- function(data) apply(data,2,function(x)
                              x[sample(1:length(x),size=length(x),replace=T)])
# Функция создания перевыборки с использованием нуль-модели r2dtable
boot1 <- function(data) r2dtable(1, rowSums(data), colSums(data))[[1]]
# Построение графика пучков линий бутстрепа
plot(Spec.pca$CA$eig, pch=16, type="b",ylab="Собственные значения",
                                    xlab="Номер компоненты", xlim=c(1,8))
    for(i in 1:100) lines(rda(boot2(Bentos),scale=FALSE)$CA$eig, col="red")
    for(i in 1:100) lines(rda(boot(Bentos),scale=FALSE)$CA$eig, col="grey")
    for(i in 1:100) lines(rda(boot1(Bentos),scale=FALSE)$CA$eig, col="green")
  points(Spec.pca$CA$eig, type="b", lwd=2)  #  Линия по эмпирической выборке
#  Построение ординационных диаграмм
plot(Spec.pca, dis = "sites", type = "n")    # По объектам (участкам реки)
ordipointlabel(Spec.pca, display = "sites", font=4,  pch=21,  bg="grey")
plot(Spec.pca,  dis = "sp", type = "n")    # По переменным (видам макрозообентоса)
stems <- colSums(Bentos)  ; ordilabel(Spec.pca, dis = "sp", priority=stems)
```

6.3. Сравнение результатов различных моделей ординации

"Идеальный" метод ординации должен обладать рядом следующих свойств:

- близость объектов (например, композиций видов и местообитаний) в ординационном пространстве должна быть тесно связана с их экологическим подобием;
- главные градиенты изменчивости должны отображаться без искажений, а принятый масштаб осей определяться разнообразием сообществ;
- если в изучаемой системе объективно существует возможность группировки, то это находит свое отражение при ординации, однако не существующие псевдо-кластеры на диаграмме не обнаруживаются;
- не чувствительность к "шуму", однако "сигнал" и "шум" надежно отличаемы;
- результативность и воспроизводимость в различных условиях, т. е. одинаковая работоспособность для разреженных и заполненных матриц, а также наборов данных различной размерности;
- математическая элегантность, доступная и легкая для понимания пользователей.

Метод PCA, тесно связанный со свойствами ковариационной матрицы, не всегда отвечает этим условиям, поэтому была разработана целая гамма ординационных методов (Legendre, Legendre, 1998), основанных на различных алгоритмах обработки данных. *Анализ главных координат* (PCoA), или метод метрического многомерного шкалирования (Айвазян и др., 1989) в определенной степени абстрагируется от исходной таблицы данных мониторинга и может оперировать с произвольной матрицей расстояний **D**, где d_{ij} – мера дистанции между каждой парой местообитаний i и j, $i, j = 1, 2, \ldots, n$.

Другим перспективным методом ординации, находящим все большее применение в экологии, является алгоритм *неметрического многомерного шкалирования* (NMDS, nonmetric multidimensional scaling – Дэйвисон, 1988), также использующий произвольную матрицу дистанций **D**. Считается, что этот метод дает наиболее адекватные результаты, особенно для больших биогеографических матриц с сильными шумами (Minchin, 1987). Его главным преимуществом является то, что он не требует от исходных данных никаких априорных предположений о характере статистического распределения.

В общем случае, задача многомерного шкалирования состоит в том, чтобы создать такой p-мерный ($p = 2$ или 3) "образ" наших объектов (видов и местообитаний), в котором взаимные попарные расстояния оказались бы наименее искажены по сравнению с исходным состоянием **D**. Главные оси геометрической метафоры данных в пространстве ме́ньшей размерности обычно находятся путем минимизации критерия "стресса":

$\Delta = \sum_{i,j=1}^{n} d_{ij}^{\alpha} \left| \hat{d}_{ij} - d_{ij} \right|^{\beta}$, где d_{ij} и $\hat{d}_{ij}$ – расстояния между объектами i и j в исходном и редуцированном пространствах, α и β – задаваемые коэффициенты.

Отличие между метрической и неметрической модификациями заключается в том, что поиск решения PCoA осуществляется на множестве линейных функций (с точностью до ортогональных преобразований) и основан на операциях с собственными числами и собственными векторами (Айвазян и др., 1989, с. 439). В методе NMS выполняется последовательность итераций для минимизации критерия Δ, оценивающего степень сходства между исходной и моделируемой матрицами расстояний. Детальное описание обоих методов широко представлено в доступной методической литературе.

Для реализации метода PCoA предварительно рассчитаем матрицу расстояний D размером 159×159 в пространстве 22 видов травянистой растительности между каждой парой геоботанических описаний в дельте р. Волга (пример [П3]). Выбор конкретной расчетной формулы для метрики дистанции выполним, сравнив коэффициенты ранговой корреляции между двумя матрицами расстояний: по видовой структуре на основе испытуемого индекса и на основе факторов среды (в конкретном примере – ионный состав почвы и ее увлажненность). Наилучшим индексом по этому тесту оказался коэффициент Брея-Кёртиса, опередивший евклидово и манхеттенское расстояние, а также другие традиционные экологические меры.

Как и в методе PCA, в случае многомерного шкалирования исследуются зависимости не только между объектами (Q-анализ), но и между переменными (R-анализ). На совмещенном биплоте проекции объектов и переменных связываются между собой с использованием коэффициентов корреляции или взвешенных средних. Ординационная диаграмма видов на рис. 6.8, построенная для нашего примера на основе PCoA, выгодно отличается от биплота PCA (см. рис.6.4) более высокой разрешающей способностью, что позволяет провести важное в предметном отношении выделение структурных ассоциаций в изучаемом растительном сообществе.

С использованием линейных аддитивных моделей можно найти регрессионную зависимость между координатами построенной ординации и факторами окружающей среды. Функция ordisurf() пакета vegan выполняет автоматическое кусочно-линейное сглаживание (thin plate spline) результатов моделирования и строит изономы поверхности распределения рассматриваемого фактора в пространстве двух шкал. Это позволяет оценить, какое место занимает каждый вид на градиенте внешних условий: в нашем примере на рис. 6.4 – по отношению к увлажненности почвы и уровню ее минерализации.

Для другого примера [П2], ординации 13 участков рек Сок-Байтуган и 129 видов макрозообентоса, используем метод NMDS. Здесь также актуальна проблема выбора наилучшей метрики расстояния. Кроме того, в этом методе расчет шкал принято повторять несколько раз, поскольку, из-за нелинейных отношений между объектами в исходном и редуцированном пространствах, оптимизационная процедура может легко "застрять" в локальном минимуме функции стресса. Функция metaMDS(…) пакета vegan на этом примере выполнила в комплексе действия, приведшие к следующему результату:

° провела выбор лучшей ординационной метрики, какой оказалось расстояние Брея-Кёртиса;

° после 20 серий расчетов нашла конфигурацию шкал с минимальным стрессом $\Delta = \sqrt{\sum_{i \neq j} [\theta(d_{ij}) - \hat{d}_{ij}]^2 / \sum_{i \neq j} \hat{d}_{ij}^2} = 0.0574$ (или с максимальной корреляцией, основанной на стрессе $R^2 = 1 - \Delta^2 = 0.997$);

° повернула график относительно начала координат, чтобы наибольшая дисперсия меток участков была на первой оси, а масштаб шкал определила так, чтобы каждая их единица соответствовала делению облака точек примерно пополам;

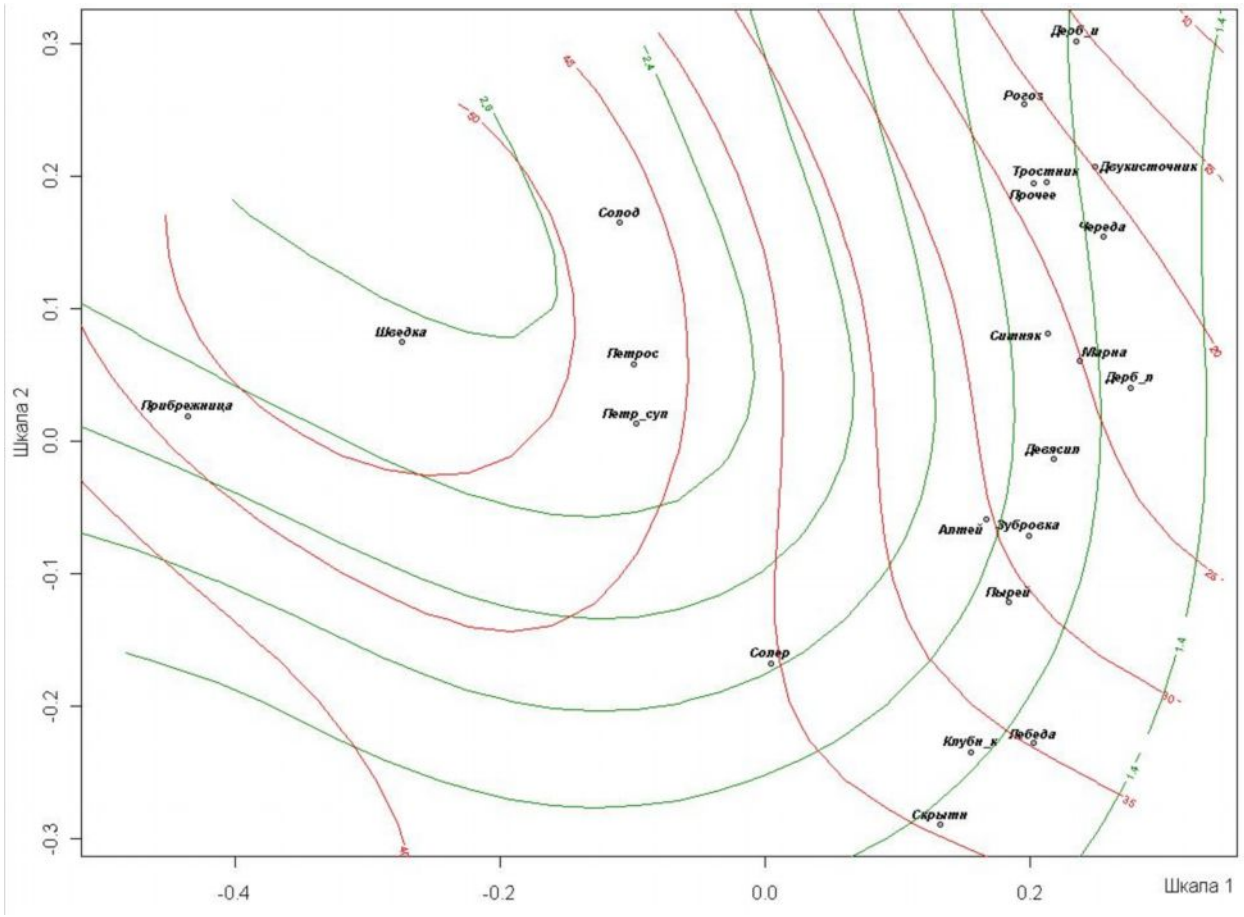


Рис. 6.8. Ординация видового состава растительности в дельте р.Волга методом PCoA (зеленым цветом представлены изономы высот над уровнем моря, красным – суммарного содержания ионов)

- для каждого проецируемого объекта рассчитала вклад в общий стресс Δ (на рис. 6.9 диаметр кружков пропорционален неопределенности координат участков при ординации);
- нашла координаты меток видов, как взвешенные средние от координат участков.

В предыдущем разделе на рис. 6.6 мы получили аналогичную ординацию участков рек Сок-Байтуган методом PCA. Чтобы сравнить, насколько отличаются две диаграммы, мы должны привести их к единой *форме*, т.е. к стандартной геометрической конфигурации, которая независима от положения, масштаба и поворота координатных осей. Один из способов сравнения графиков заключается в "прокрустовом" преобразовании геометрии точек. Для этого проводится ряд последовательных итераций подгонки масштаба осей:

- сравниваемые точки помещаются внутри единичного круга $|\delta| = 1$;
- вся структура вращается относительно центра координат;
- находится минимум суммы квадратов расстояний между двумя ординациями $m^2 = [\mathbf{x}_1 - T(\mathbf{x}_2)]^2 \rightarrow \min$ или максимум прокрустовой корреляции $r = \sqrt{1 - m^2}$.

При таком изометрическом выравнивании относительные расстояния между точками каждой диаграммы не меняются, но структуры приобретают общий центр тяжести, примерно одинаковый размер и ориентацию. Диаграмма (рис. 6.10) прокрустовой суперпозиции ординаций показывает, что неметрическое шкалирование в меньшей мере выделяет различие между участками р.Байтуган и верхней части р.Сок, чем это делает метод главных компонент, но контрастнее оценивает своеобразие устьевых участков С-12 и С-13.

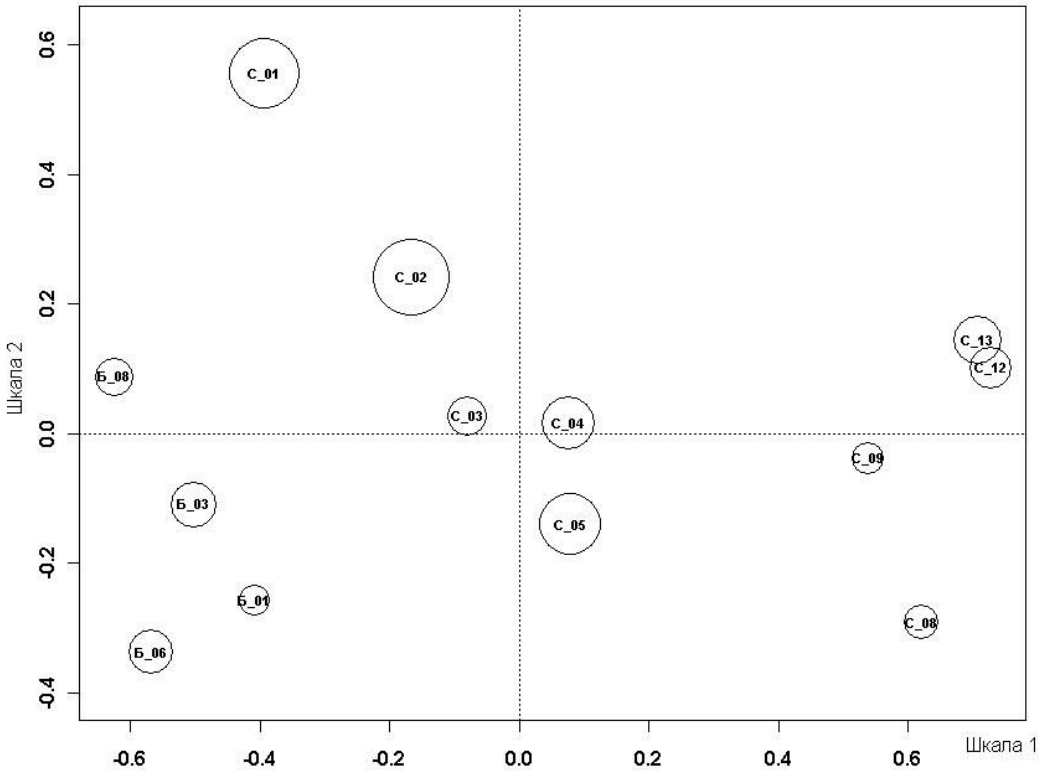


Рис. 6.9. Ординация участков рек Сок-Байтуган методом неметрического многомерного шкалирования

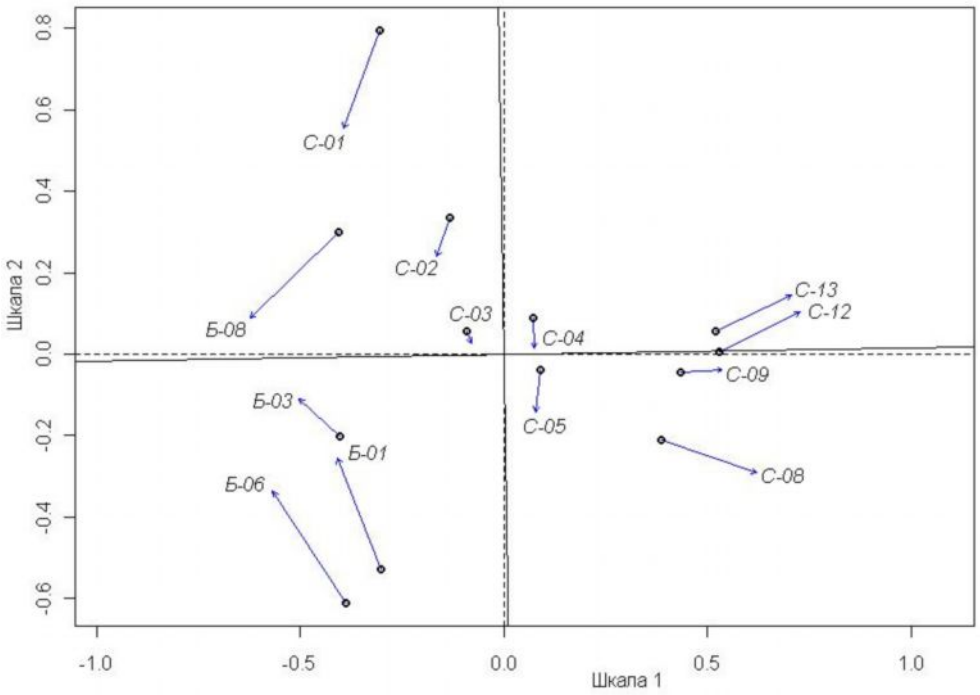


Рис. 6.10. Диаграмма суперпозиции ординаций участков рек Сок-Байтуган, полученных методами PCA (кружки) и NMDS (стрелки с меткой), при прокрустовом расстоянии между ними $m^2 = 0.552$

Непрямой ординационный анализ только одной матрицы **X** дает отображение в ортогональной системе координат частных структурных особенностей экосистемы в форме графических проекций видов и их местообитаний. В экологических исследованиях это – пассивный путь, не дающий непосредственной оценки зависимости видовой структуры сообществ от факторов окружающей среды, что требует от исследователя апостериорной интерпретации, основанной часто на субъективных предположениях.

Каноническая ординация позволяет выполнить совместную обработку двух или более наборов данных и проверить статистические гипотезы о значимости как внутренних взаимодействий, так и о влиянии внешних факторов. Концептуально канонический анализ позиционируется как расширение регрессионного анализа при моделировании многомерного отклика данных: $\mathbf{Y} = f(\mathbf{X})$,

где $\mathbf{Y}$ – матрица $m \times n$, содержащая центрированные значения обилия y_{ij} по m видам (строки) и n местообитаниям (столбцы); $\mathbf{X}$ – матрица $q \times n$, в которой j-я строка содержит центрированные значения фактора среды x_{kj}.

Наиболее часто используются две математические процедуры канонической ординации, известные под аббревиатурами RDA и CCA. Анализ избыточности (RDA, redundancy analysis – Rao, 1964; Legendre, Legendre, 1998) основан на общей линейной модели регрессии и является канонической формой метода главных компонент PCA. Технически эта процедура выглядит следующим образом:

° рассчитывается регрессия каждой переменной $\mathbf{y}_i$ на таблицу факторов **X,** в результате чего вычисляются предикторные значения $\hat{y}_{ij}$ и остатки y_{res};

° выполняется PCA-анализ матрицы $\hat{\mathbf{Y}}$ и вычисляются канонические собственные значения и собственные векторы **U**;

° матрица **U** используется для расчета компонентов ординации двух типов: общих взвешенных сумм счетов объектов (или scores $\mathbf{YU}$) и сумм счетов, обусловленных линейными комбинациями влияющих факторов среды (или site constraints $\hat{\mathbf{Y}}\mathbf{U}$);

° PCA-анализ выполняется также с матрицей остатков множественной регрессии $\mathbf{Y}_{res}$ и рассчитываются оси главных компонент, независимые (unconstrained) от влияния внешних факторов.

Задачей *канонического анализа соответствий* (CCA или canonical correspondence analysis – ter Braak, 1986) является определение таких коэффициентов для видов $\mathbf{a} = (a_i)$, $i = 1, ..., m$, и для факторов среды $\mathbf{c} = (c_k)$, $k = 1, ..., q$, которые делают максимальной корреляцию между $\mathbf{z^*} = \mathbf{Y'a}$ и $\mathbf{z} = \mathbf{X'c}$. В целом, CCA является модификацией RDA, выполняющей "взвешивание" ординационных компонент пропорционально их вкладу в статистику χ^2, оценивающую расстояния между объектами в многомерном пространстве. CCA также основан на аппроксимации данных многомерной гауссовой регрессией, но в экологических приложениях придает существенно бóльшее значение редко встречающимся видам, чем это делает RDA.

Диаграммы ординации, полученные CCA или RDA, как и в случае PCA, представляют собой совмещенный график, где точки видов находятся в центрах тяжести распределения их популяционной плотности, а координаты местообитаний определяются взвешенной комбинацией обилия характерных для них видов. Однако в каноническом анализе характер распределения $\{y_{ki}\}$ дополнительно описывается с помощью гауссовой модели отклика, в которой объясняющая переменная является линейной комбинацией факторов окружающей среды $\{c_k \cdot x_{ji}\}$. Соответственно, результирующая диаграмма ("триплот" – см. рис. 6.11) отражает не только изменчивость видового состава относительно двух осей проецирования F_1-F_2, но и статистические связи между видами и каждой независимой переменной $\{x_{ji}\}$. Для этого из центра координат стрелками проводятся дополнительные оси физических градиентов, ориентация которых зависит от значений канонических собственных векторов. Переменные среды, обозначаемые длинными стрелками, сильнее связаны с осями ординации F_1-F_2, чем факторы, обозначаемые короткими стрелками и, следовательно, более значимо определяют изменчивость структуры экосистемы.

Проекция точки вида на каждую стрелку показывает экологический оптимум (точнее, центр тяжести распределения обилия) этого вида относительно анализируемого физического фактора. Если виды представить стрелками, исходящими из начала координат, то косинус угла между стрелкой вида и стрелкой фактора среды

приблизительно равен коэффициенту корреляции между ними. Кроме того, стрелки, имеющие примерно одно направление, говорят о значимой положительной корреляции, стрелки, пересекающиеся под прямым углом, говорят о корреляции, близкой к нулю, а стрелки, имеющие примерно противоположное направление, указывают на высокую отрицательную корреляцию.

Пример ординационной CCA-диаграммы с наложением дополнительных осей объясняющих переменных, представленный на рис. 6.11, содержит в концентрированной форме богатый материал для различных экологических интерпретаций и умозаключений. Нетрудно заметить, например, что ось 1 ординации связана с комплексом гидрологических и гидрохимических показателей, определяющих продольную изменчивость условий среды в реках от их истоков к устью, а ось 2 во многом определяется площадью аккумуляции поверхностного стока, органическим загрязнением и содержанием растворенного кислорода.

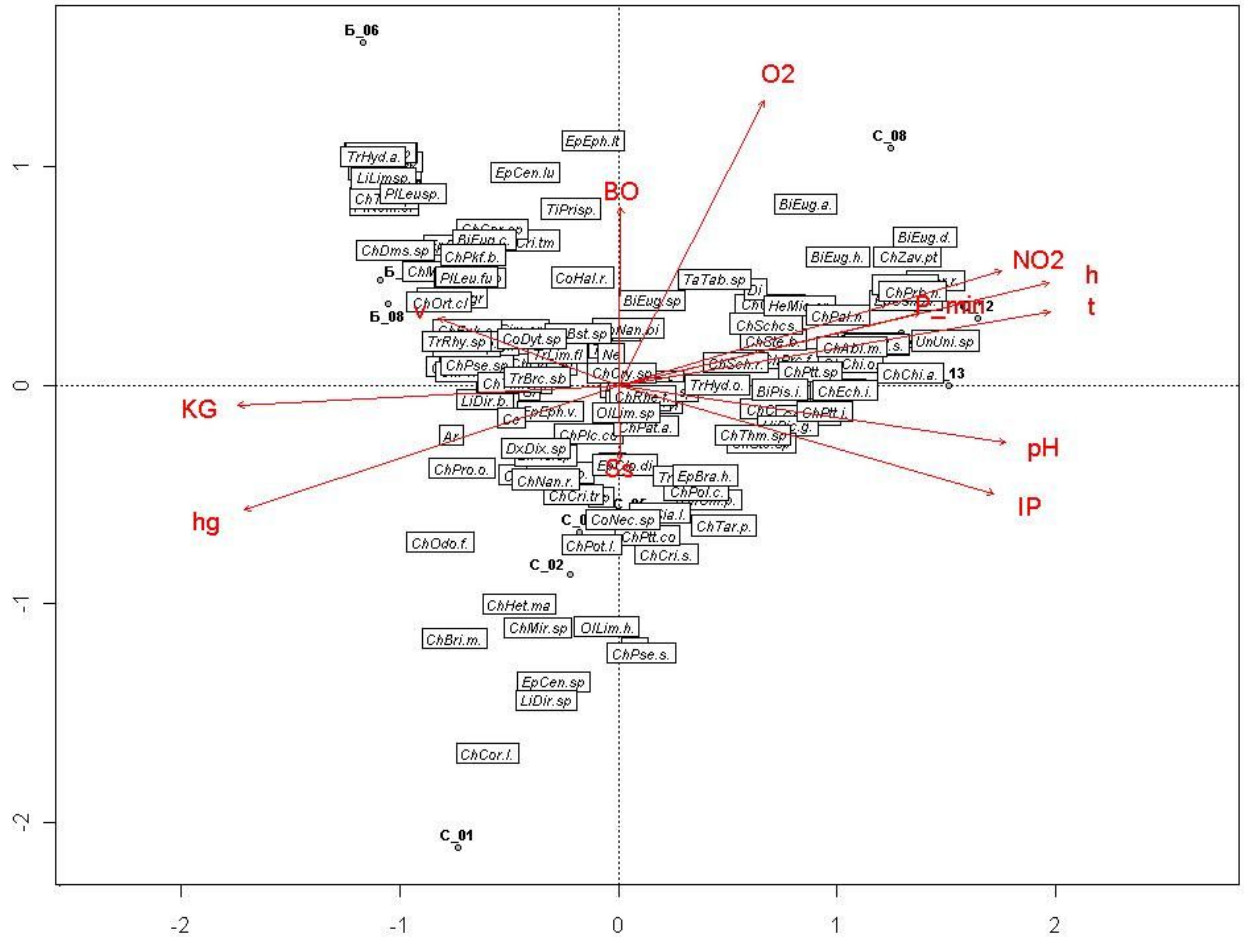


Рис. 6.11. Совмещенная диаграмма (триплот) канонического анализа соответствий CCA для ординации видовой структуры макрозообентоса на участках рек Байтуган-Сок под воздействием факторов среды (**v** – скорость течения, **t** – температура, **h** – глубина в местах отбора проб, **pH** – активная реакция среды, **BO** – бихроматная окисляемость, $\mathbf{NO_2}$ – нитритный азот, $\mathbf{O_2}$ – кислород, $\mathbf{P_{min}}$ – фосфор минеральный, **KG** – каменистость грунта, **IP** – заиленность грунта, **hg** – высота над уровнем моря, **Ss** – площадь водосбора)

Используем, однако, аппарат проверки статистических гипотез на основе рандомизации, чтобы оценить формальную достоверность предполагаемых зависимостей. Зададимся вначале вопросом о значимости коэффициентов корреляции каждого отдельного фактора среды с распределением видов по участкам $r^2 = 1 - SS_{res}/SS_{tot}$, где SS_{res} и SS_{tot} – остаточная и полная суммы квадратов. После 1000 операций случайного перемешивания векторов матрицы **X** пермутационный тест показал, что 3 фактора из 12

оказались статистически незначимыми: скорость течения **v** (p = 0.375), площадь водосбора **Ss** (p = 0.866) и бихроматная окисляемость **BO** (p = 0.416).

Однако каноническая ординация принимает во внимание не отдельные факторы среды, а такие их линейные комбинации, которые дают наименьшую общую остаточную дисперсию. И здесь возникает традиционная задача поиска наилучшей модели регрессии, в которой были бы исключены неинформативные компоненты, провоцирующие неустойчивость решения. Это особенно важно в условиях высокой корреляции влияющих переменных, что легко установить для нашего примера визуально или с использованием процедуры оценки коэффициентов инфляции дисперсии (VIF – variance inflation factor).

Для оценки качества моделей-претендентов при их селекции могут быть использованы следующие статистики, обсуждаемые нами ранее в главах 3 и 4:

- коэффициент детерминации множественной регрессии R^2 и приведенный к числу степеней свободы коэффициент детерминации $R^2_{adj} = 1 - \frac{n-1}{n-q-1}(1-R^2)$;
- статистика, подобная информационному критерию Акаике AIC;
- псевдо-F-статистика, оцениваемая как отношение суммы квадратов $SS(\hat{\mathbf{Y}})$ к остаточной сумме квадратов $SS(\mathbf{Y}_{res})$ с учетом числа степеней свободы.

Рассмотрим три модели-претендента с характеристиками, представленными в табл. 6.3:

- модель 1 от всех 9 факторов среды со значимым коэффициентом корреляции r^2;
- модель 2, полученная шаговыми процедурами с учетом итогов рандомизации;
- модель 3, полученная стандартной шаговой процедурой step(), ищущей модель с минимальным значением AIC.

Таблица 6.3. Статистические характеристики трех моделей RDA по данным гидробиологической съемки на реках Сок-Байтуган (обозначения факторов см. на рис. 6.11)

Число переменных	Коэффициент детерминации		Псевдо-критерий AIC	Псевдо-F-статистика	
	обычный	приведенный		F	Pr(>F)
Модель 1: Y ~ **hg** + **t** + **h** + **O_2** + **P_{min}** + **NO_2** + **pH** + **KG** + **IP**					
9	0.835	0.34	-8.97	1.68	0.018
Модель 2: Y ~ **t** + **IP** + **Ss**					
3	0.51	0.346	-12.9	3.12	0.001
Модель 3: Y ~ **t**					
1	0.343	0.282	-13.07	3.76	0.001

Расчет моделей проводился с использованием двух функций: forward.sel(), максимизирующей приведенный коэффициент детерминации R^2_{adj}, и ordistep(), находящей минимум AIC. При этом ставилось дополнительное условие, что все включенные переменные должны быть статистически значимыми, а все исключенные переменные – незначимыми в соответствии с пермутационным тестом, оценивающим частные F-статистики. На этом фоне модель 1 видится переопределенной, а модель 3 – незаслуженно урезанной, что явилось следствием некоторой математической условности подсчета критерия Акаике при RDA-анализе. Модель 2 представляется нам наиболее сбалансированной.

Дополнительную информацию для анализа моделей предоставляет дисперсионный анализ anova.cca, оценивающий по частным F-критериям статистическую значимость термов модели и осей канонических ординаций с использованием рандомизации.

К разделу 6.3:

```
# 1. Анализ главных координат  - геоботанические описания в дельте р. Волга
# Загрузка таблицы видов из предварительно подготовленного двоичного файла
library(vegan)  ; load(file="Fito_Full.RData");  Species <- Species[,-8]
# Оценка коэффициентов корреляции Спирмена для различных метрик расстояния
Env <- Fito_Full[,27:36] ; rankindex(scale(Env), Species, c("euc","man","bray","jac","kul"))
spe.bray <- vegdist(Species)                # PCoA по расстоянию Брея-Кертиса
```

```
spe.b.pcoa <- cmdscale(spe.bray, k=(ncol(Species)), eig=TRUE)
spe.b.wa <- wascores(spe.b.pcoa$points[,1:2], Species)
# График взвешенных средних проекций видов
ordiplot(scores(spe.b.pcoa)[,c(1,2)], type="n", ylim=c(-0.4,0.4))
abline(h=0, lty=3); abline(v=0, lty=3)
ordipointlabel(spe.b.wa, display = "sp", font=4,  pch=21, cex=0.7, bg="grey")
#  Формирование и вывод на биплот изоном факторов среды  H и Sum_all
with(Env, ordisurf(spe.b.pcoa, H, add = TRUE, col = "green4"))
with(Env, ordisurf(spe.b.pcoa, Sum_all, add = TRUE, col = "red"))
# 2. Неметрическое многомерное шкалирование (NMDS)  - макрозообентос участков р. Сок
library(xlsReadWrite)  #  Загрузка данных из файла Excel и нормирование
A <- read.xls("Сок1.xls", sheet = 1, rowNames=TRUE)
A[is.na(A)] <- 0 ; A.norm <- decostand(t(A), "normalize")
#     NMDS-анализ  и выделение вкладов участков в общий стресс
ben.mds <- metaMDS(A.norm, trace = FALSE)  ;  gof <- goodness(ben.mds)
#  Вывод графика участков с метками, диаметр которых пропорционален неопределенности
plot(ben.mds, type = "t", disp="sites", font=2) ; points(ben.mds,
disp="sites",cex=6*gof/mean(gof))
#  Сравнение диаграмм  NMDS и PCA  с использованием прокрустового преобразования
pro <- procrustes(ben.mds, rda(A.norm, scale=FALSE))
plot(pro) ; points(pro, display= "rotated") ; text(pro, display= "target")
# 3. Каноническая ординация
# Стандартизация матрицы факторов среды
Env <- read.xls("Сок1.xls", sheet = 2, rowNames=TRUE)
Env.stand<-decostand(Env,"standardize")
#     Оценка угловых коэффициентов и корреляций факторов с осями
bentos.ca <- cca(A.norm) ; (ef <- envfit(bentos.ca, Env.stand, permutations = 999))
plot(bentos.ca,  type="n")          # Отрисовка ординационного триплота
ordipointlabel(bentos.ca, display = "site", font=2,  pch=21, cex=0.7, bg="grey")
ordilabel(bentos.ca, dis = "sp", cex=0.7, font=3 ,add = TRUE)
plot(ef, cex=1.2,  col = "red", lwd=2)
#     Исследование модели 1
ord9.rda <- rda(formula = A.norm ~ hg  + t + h + O2 + P_min + NO2 + pH + KG + IP,
                                                    data = Env.stand)
RsquareAdj(ord9.rda)$adj.r.squared   ; RsquareAdj(ord9.rda)$r.squared
anova(ord9.rda, step=1000)
anova(ord9.rda,by = "terms", step=1000) ; anova(ord9.rda,by = "axis", step=1000)
#     Селекция моделей тремя различными шаговыми процедурами
library(packfor); forward.sel(A.norm, Env.stand)
ord.all <- rda(A.norm ~., data = Env.stand) ; ord.0   <- rda(A.norm ~ 1, data=Env.stand)
ordistep(ord.0, scope = formula(ord.all), pstep = 1000)
step(ord.0, scope = formula(ord.all), test = "perm", steps = 1000)
ord3.rda <- rda(formula = A.norm ~  t + IP + Ss, data = Env.stand)  #   Модель 2
ord1.rda <- rda(formula = A.norm ~  t, data = Env.stand)                  #   Модель 3
#  Для моделей 2 и 3 выполняется расчет статистик и anova, как и для модели 1
```

6.4. Деревья классификации и регрессии

Метод построения деревьев классификации и регрессии (Classification and Regression Tree, CART – Breiman at al., 1984; McCune at al., 2002; Шитиков и др., 2005) предлагает новую и весьма перспективную альтернативу выявления различий между группами, одновременно выполняя функции прогнозирования. Классификационные модели деревьев рекурсивно делят набор данных на подмножества, являющиеся все более и более гомогенными относительно определенных признаков. При этом выполняется классификация иерархического типа и формируется ассоциативный дихотомический ключ, дающий возможность выполнять распознавание объектов неизвестных выборок. Отличие классификационных и регрессионных моделей заключается в том, что в деревьях первого типа зависимая переменная измеряется в категориальных шкалах (например, характер древостоя), когда как деревья регрессии предсказывают непрерывные значения отклика (например, средний диаметр древостоя).

По своей сути деревья CART используют принцип "наивной" классификации (naive approach), поскольку исходят из предположения о взаимной независимости признаков. Поэтому модели классификационных деревьев статистически наиболее работоспособны, когда комплекс анализируемых переменных не является аддитивным или мультипликативным. Из-за своей рекурсивной природы этот метод особенно применим в случаях, когда имеется регулярная внутренняя множественная альтернатива в исходной комбинации переменных, связанная с самим процессом группировки.

Отметим несомненные преимущества моделей распознавания на основе деревьев:

1. Деревья классификации и регрессии дают возможность извлекать правила из базы данных на естественном языке. Поэтому результат работы алгоритмов CART очень легко интерпретировать визуально, что делает их особенно полезными для исследовательского анализа данных.

2. Результаты анализа содержат полную оценку того, насколько различаются между собой выделенные группы и за счет каких переменных обуславливается это отличие, а также прогнозирующую модель, с помощью которой можно предсказывать класс неизвестного объекта. Как любая модель, основанная на рекурсии, деревья позволяют вычленить множество визуально очевидных связей и отношений между переменными (некоторые из них могут иметь вполне очерченный экологический смысл), что не всегда является возможным при работе с обычными статистическими линейными моделями.

3. Как любой непараметрический метод, построение деревьев не зависит от закономерностей статистического распределения данных. Деревья CART также разумно малочувствительны к пропускам или аномальным выбросам значений, особенно при использовании уже настроенной модели для "экзамена" новых объектов.

4. Деревья позволяют хранить информацию о данных в компактной форме, т.е. вместо обширных таблиц данных мы можем хранить дерево решений, которое содержит в концентрированной форме точное описание объектов.

Деревья классификации и регрессии представляют собой последовательные иерархические структуры, состоящие из узлов, которые содержат правила, т.е. логические конструкции вида "если …, то …". Корневой узел дерева связан с граничным значением одной из переменных исходной таблицы данных, которое делит все множество объектов на две группы (для бинарного случая). От каждого последующего узла-родителя к узлам-потомкам также может отходить по две ветви, в свою очередь связанные с граничными значениями других наиболее подходящих переменных и определяющие правила дальнейшего разделения {splitting criterion} на группы. Конечными узлами дерева являются "листья", соответствующие найденным решениям и объединяющие всё множество объектов классифицируемой выборки. Общее правило выбора опорного значения для каждого узла построенного дерева можно сформулировать следующим образом: «выбранный признак должен разбить множество $\mathbf{X}^*$ так, чтобы получаемые в итоге подмножества $\mathbf{X}^*_\mathbf{k}$, $k = 1, 2, \ldots, p$, состояли из объектов, принадлежащих к одному классу, или были максимально приближены к этому».

Построение деревьев осуществляется, как правило, с использованием "жадных" алгоритмов, стремящихся, не считаясь ни с чем, построить модель с максимально возможным числом сомнительных и посторонних включений. Естественно, чем обширнее и кустистее дерево, тем лучшие результаты тестирования оно показывает на обучающей выборке, но не столь успешно выполняет экзамен незнакомых примеров. Поэтому построенная модель должно быть еще и оптимальной по размерам, т.е. содержать информацию, улучшающую качество распознавания, и игнорировать ту информацию, которая его не улучшает. Для этого обычно проводят "обрезание" дерева (tree pruning) – отсечение ветвей там, где эта процедура не приводит к серьезному возрастанию ошибки.

Лишь в редких случаях удается подобрать объективный внутренний критерий, приводящий к хорошему компромиссу между безошибочностью и компактностью, поэтому стандартный механизм оптимизации деревьев основан на кросс-проверке. Для

этого обучающая выборка разделяется, например, на 10 равных частей: 9 частей используется для построения и последующего обрезания дерева, а оставшаяся часть играет роль внешнего дополнения для независимого экзамена. После многократного повторения этой процедуры из некоторого набора деревьев-претендентов, у которых практически допустимый разброс остальных параметров качества, выбирается дерево, показавшее наилучший результат при кросс-проверке.

В качестве примера [П4] рассмотрим построение дерева CART, прогнозирующего зоны обитания М популяций красной полевки, которые расположены на различном расстоянии от Байкальского ЦБК и принимают следующие значения: 1 – до 5 км, 2 – от 15 до 50 км, 3 – свыше 50 км. В качестве опорных переменных будем использовать морфометрические показатели: массу внутренних органов животных, а также длину и массу тела, взаимосвязь между которыми анализировалась ранее в разделе 4.6.

В общем случае может быть использовано несколько алгоритмов построения деревьев на основе различных схем и критериев оптимизации. Функция rpart(…) из одноименного пакета выполняет рекурсивный выбор для каждого следующего узла таких разделяющих значений, которые приводят к минимальной сумме квадратов внутригрупповых отклонений D_t для всех t узлов дерева. При решении задачи классификации *девианс* D_i имеет смысл "числа посторонних включений" и оценивается по формулам расчета энтропии или индекса Джини. Для дерева, представленного на рис. 6.12, начальное разбиение дало две группы особей: 19 полевок 1-го класса (энтропия $D_1 = 0$) с массой тела **W** > 44 г и общую группу из остальных 259 не столь упитанных мышей разных классов ($D_2 = 0.93$), которая подвергается дальнейшему разбиению.

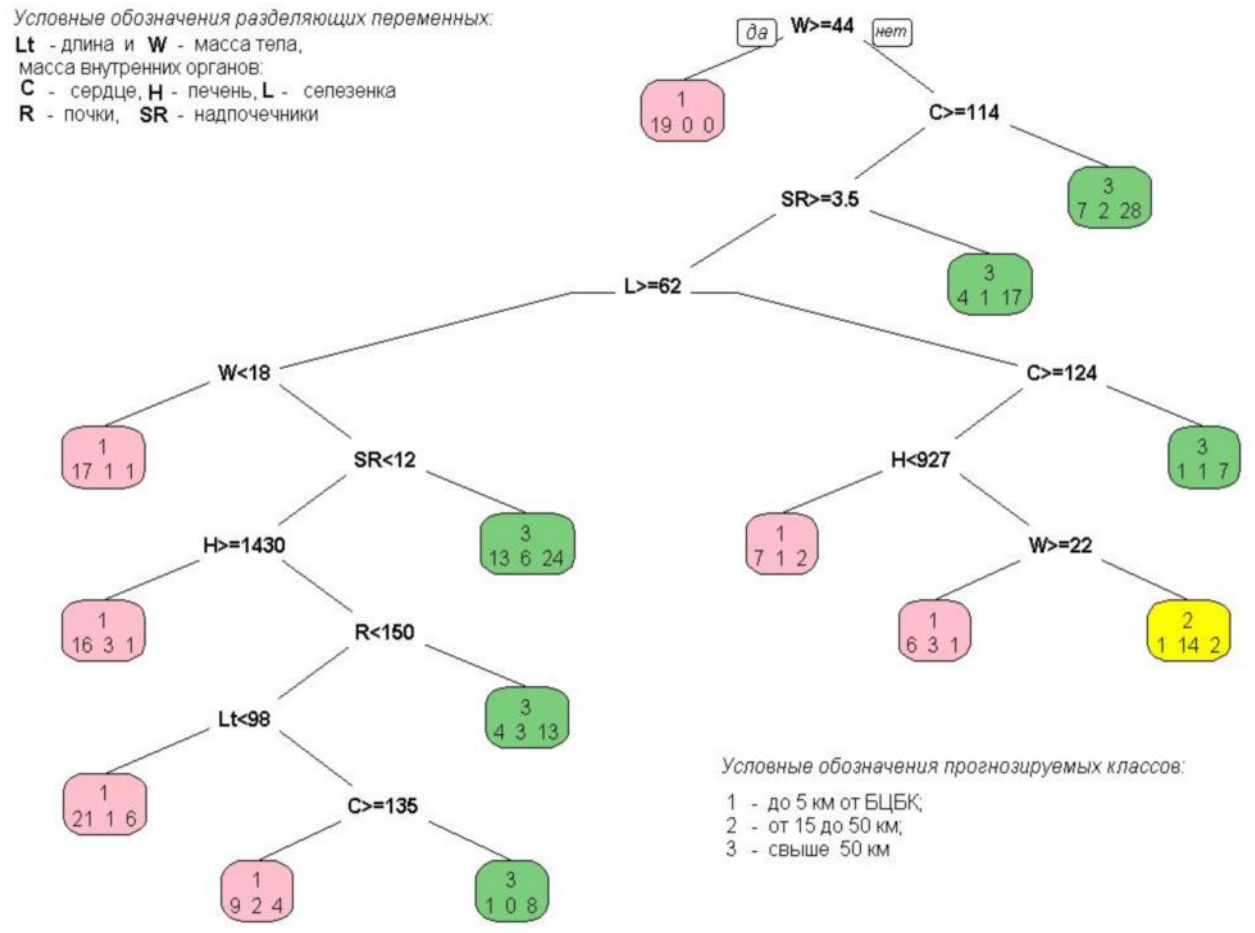


Рис. 6.12. Дерево **rpart** для распознавания местоположения популяций красной полевки (в "листьях" указаны номер прогнозируемого класса и распределение числа особей обучающей выборки по их фактической групповой принадлежности)

Оценка качества построенного дерева **T** в ходе его оптимизации проводилась с использованием совокупности различных критериев:

- критерия стоимости сложности (*cost complexity*), включающего штрафной множитель λ за каждую неотсечённую ветвь – $CC(\mathbf{T}) = \sum_t D_t + \lambda t$;
- девианса D_0 для нулевого дерева (т.е. оценки изменчивости в исходных данных);
- относительного параметра стоимости сложности $C_p = \lambda / D_0$;
- относительной ошибки обучения (REL_{er}) для дерева из t узлов $\sum_t D_t / D_0$;
- ошибки кросс-проверки (CV_{er}) с разбиением на 10 блоков, также отнесенной к девиансу нуль-дерева D_0; CV_{er}, как правило, больше, чем REL_{er};
- стандартного отклонения (*SE*) для ошибка кросс-проверки.

Лучшим считается дерево, состоящее из такого количества ветвей t, для которого является минимальной сумма (CV_{er} + *SE*). Для модели, классифицирующей популяции полевки по степени удаленности от БЦБК, график изменения относительной ошибки кросс-проверки, представленный на рис. 6.13, показывает, что исходное дерево из 13 узлов (см. рис. 6.12) является оптимальным и в "обрезании" не нуждается.

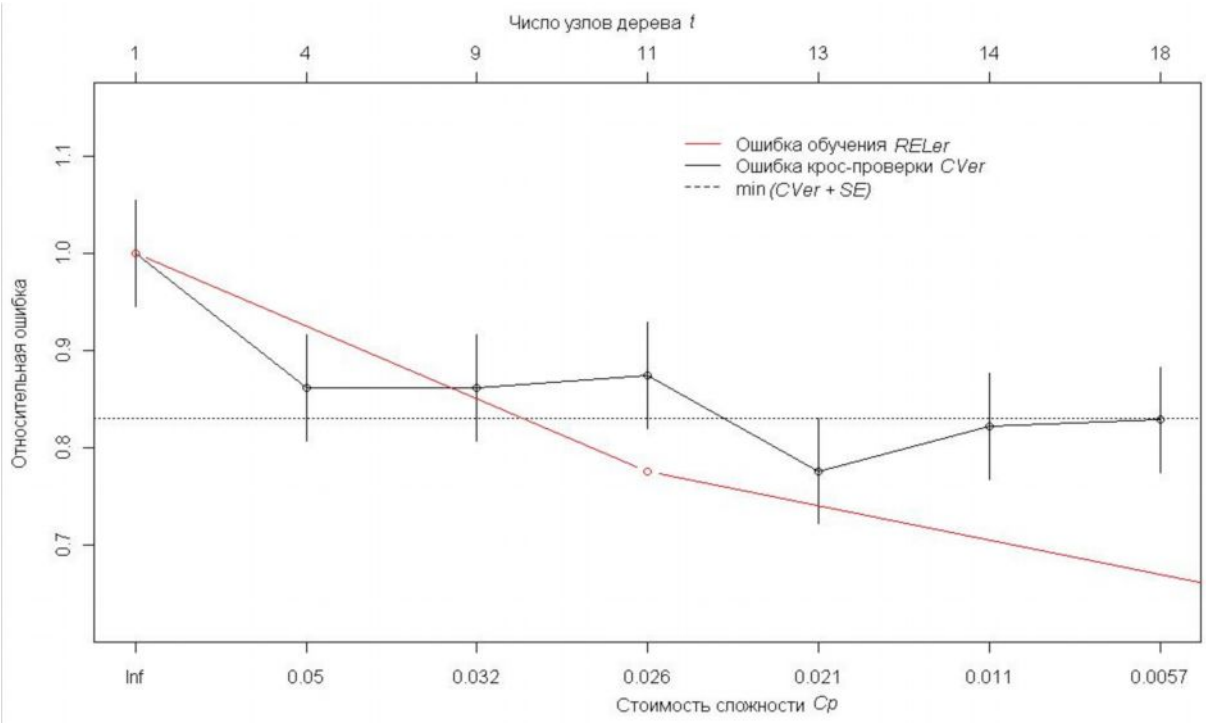


Рис. 6.13. Изменение ошибки распознавания при обучении и кросс-проверке при увеличении размеров дерева

Результаты классификации удобно представлять в виде таблицы сопряженности 6.4, где по главной диагонали расположено число правильных распознаваний. Отметим также, что при кросс-проверке доля ошибочных предсказаний существенно выше (42.4%), хотя и значительно меньше вероятности неудачи при случайном угадывании (54.7%).

Таблица 6.4. Таблица сопряженности результатов распознавания местоположения популяций красной полевки

Факти-чески	Предсказано			Итого	Число ошибок	Доля ошибок
	1	2	3			
1	**95**	1	30	126	31	24.60%
2	11	**14**	13	38	24	63.16%
3	15	2	**97**	114	17	14.91%
Итого	121	17	140	278	72	25.90%

Принципиально иной подход к построению моделей классификации основан на двух относительно новых методах ресамплинга – *баггинге* (bagging) и *бустинге* (boosting). Они используют старые идеи создания "распознающего коллектива" (Расстригин, Эренштейн, 1981), эффективность которого практически всегда оказывается значительно выше любого из его членов. Баггинг (аббревиатура от Bootstrap aggregating) на основе одно и того же метода обучения выполняет селекцию достаточно большого подмножества моделей, полученных с использованием бутстрепа исходной выборки, которые взвешиваются по уровню их эффективности и используются для экзамена с использованием голосования. Бустинг, как и баггинг, является подходом, основанным на коллективном распознавании, однако, вместо простого усреднения результатов, реализует итеративную процедуру, напоминающую градиентные алгоритмы. При этом многократно создаются взвешенные версии обучающихся выборок с весами, адаптивно подстраиваемыми на каждом шаге таким образом, чтобы попытаться в максимальной мере учесть мнение "слабых" членов коллектива, склонных на предыдущих шагах к ошибочным прогнозам.

Реализация баггинга для моделирования данных деревьями CART была выполнена Л.Бримэном (Breiman, 2001) в алгоритме под названием randomForest (см. также пакет и функцию R с тем же названием). Его использование для распознавания местоположения популяций красной полевки привело к 40.3% ошибочной классификации, что несколько лучше, чем при кросс-проверке методом rpart, и является, видимо, объективным порогом точности прогноза на использованной обучающей выборке. Однако алгоритм randomForest существенно уступает методу rpart в наглядности и объясняющей составляющей, поскольку не имеет возможности предъявить исследователю граф конкретного дерева (как на рис. 6.12), а оперирует в анализе, например, 500 деревьями, полученными бутстреп-процедурой.

Стандартный механизм проверки статистического гипотез, который предотвращает переусложнение модели, реализован в методе построения деревьев на основе "условного вывода" (conditional inference). Функция ctree(…) из пакета party принимает во внимание характер распределения независимых переменных и осуществляет на каждом шаге рекурсивного разделения данных несмещенный выбор влияющих ковариат, используя формальный тест на основе статистического критерия $c(\mathbf{t}_j, \mu_j, \Sigma_j), j = 1, ..., m$, где, μ, Σ – соответственно среднее и ковариация (Hothorn at al., 2006). Оценка статистической значимости *c*-критерия выполняется на основе пермутационного теста, в результате чего формируются компактные деревья, не требующие процедуры обрезания.

На рис. 6.14 представлено дерево "условного вывода" ctree, содержащее результаты моделирования другой категориальной переменной выборки морфометрических показателей красной полевки – пола животных. Итоги анализа не содержат сколько-нибудь неожиданных выводов: двумя статистически значимыми разделяющими переменными оказались масса надпочечников и почек.

К разделу 6.4:

```
library(xlsReadWrite) ; library(rpart) ; library(rpart.plot)
TOR <- read.xls("Ruts.xls", sheet = 1, rowNames=TRUE) ; attach (TOR)
#  Построение и отображение дерева rpart
rut.rpart <- rpart(formula = M ~ sex+W+Lt+C+R+SR+H+L, method="class", data=TOR)
prp(rut.rpart, extra=1,box.col=c("pink", "yellow", "palegreen3")[rut.rpart$frame$yval])
rut1.rpart <- rpart(M~sex+W+Lt+C+R+SR+H+L, method="class", data=TOR,
              control=rpart.control(cp=.005))  #   Снижаем порог стоимости сложности
#  График изменения относительных ошибок от числа узлов дерева
plotcp(rut1.rpart) ; with(rut1.rpart, {lines(cptable[,2]+1,cptable[,3],type="b",col="red")
   legend(locator(1),c("Ошибка обучения","Ошибка крос-проверки (CV)","min(CV ошибка)+SE"),
   lty=c(1,1,2),col=c("red","black","black"),bty="n") }) ; printcp(rut1.rpart)
table(predict(rut.rpart, type = "class"),TOR$M) # Вывод таблицы сопряженности
library(randomForest)    # Построение коллектива деревьев с использованием баггинга
rut.rf <- randomForest(as.factor(M) ~ sex+W+Lt+C+R+SR+H+L, data=TOR, importance=TRUE
```

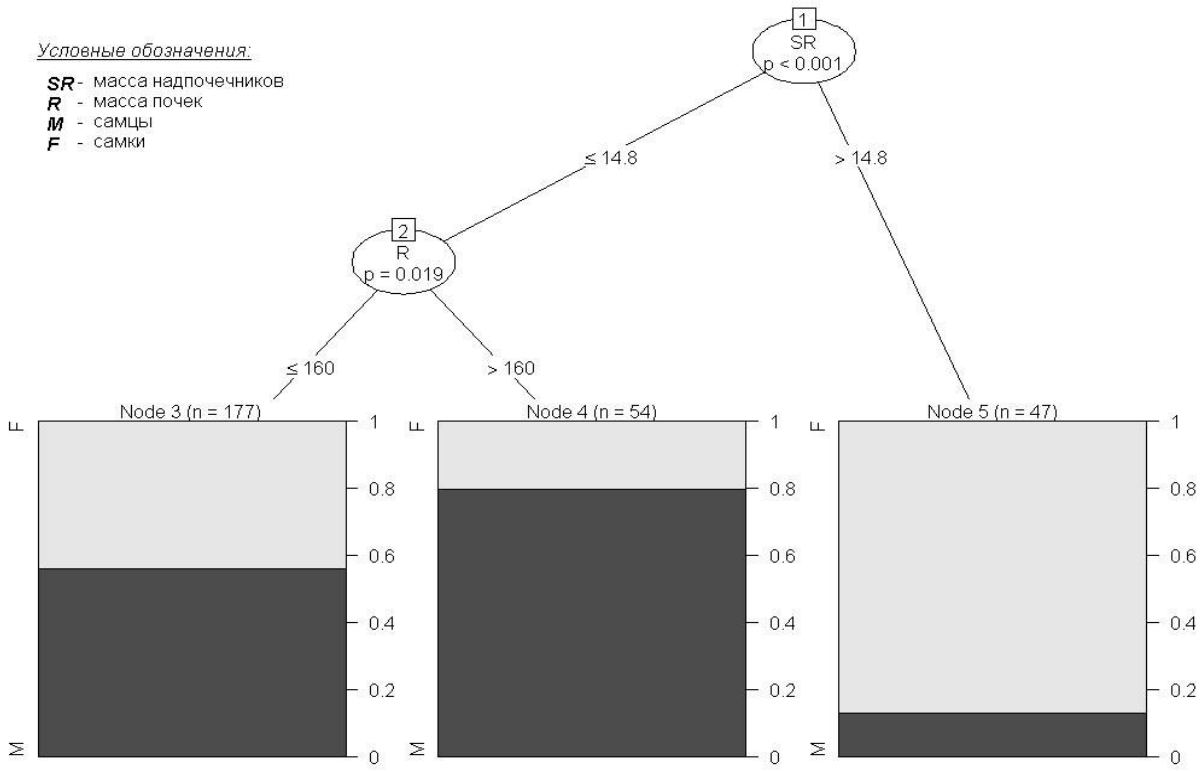


Рис. 6.14. Дерево **rpart** для распознавания местоположения популяций красной полевки

```
library(party)  # Построение дерева  на основе «условного вывода»
rut.ctree <- ctree(sex ~ W+Lt+C+R+SR+H+L, data=TOR); plot(rut.rpart) ; text(rut.rpart)
```

6.5. Деревья классификации с многомерным откликом

Развитием идеи деревьев CART, прогнозирующих одномерный отклик, являются деревья многомерной классификации и регрессии (MRT – De'Ath, 2002). Они формируются в результате рекурсивной процедуры разделения на кластеры строк двухмерной таблицы **Y**, проходящей под управлением набора внешних количественных или категориальных независимых переменных **X** (например, факторов окружающей среды). "Листьями" полученного дерева являются группы объектов (в частности, точек взятия геоботанических описаний), скомпонованные таким образом, чтобы минимизировать различия между точками в многомерном пространстве в пределах каждой совокупности.

Рассмотрим построение MRT на примере [П3] геоботанических описаний, полученных с n = 159 пробных площадок в дельте р. Волга (другие варианты обработки этого примера см. в разделах 4.5, 5.3, 5.5, 6.3). Каждая из площадок характеризуется высотой H над уровнем моря и 9-ю показателями ионного состава почвы. Процедура многомерной классификации состоит из последовательности шагов алгоритма, на каждом из которых синхронно выполняются следующие действия: (а) бинарное разделение объектов на группы, обусловленное значением одной из независимых переменных, и (б) кросс-проверка полученных результатов.

На первом шаге процедуры рассматриваются все варианты разбиения исходной выборки на две части при разных опорных значениях факторов среды и выбирается такая комбинация, которая в наибольшей мере обеспечивает экологическую гомогенность формируемых групп. Искомым критерием, минимизирующим внутригрупповые различия,

может быть, например, сумма квадратов отклонений $SS_D = \sum_{ij} (y_{ij} - \bar{y}_j)^2$, где y_{ij} – обилие вида j, обнаруженного на участке i; $\bar{y}_j$ – средние значения популяционной плотности этого вида для формируемой группы, куда включается i-й участок. Геометрически SS_D можно представить как сумму евклидовых расстояний объединяемых объектов от центра их группировки. Например, значение H = 1.2 для дерева на рис. 6.15 делит участки взятия проб на два подмножества: 130 площадок, расположенных выше 1.2 м над уровнем моря, и 29 площадок, лежащих ниже этой отметки, причем в результате этого разбиения величина SS_D уменьшается на 20%. На третьем шаге процедуры разбиения формируются два "листа" 4 (группа из 36 площадок с H > 2.3 м) и 5 (47 площадок с H от 1.8 до 2.3 м), дальнейшее разбиение которых нецелесообразно.

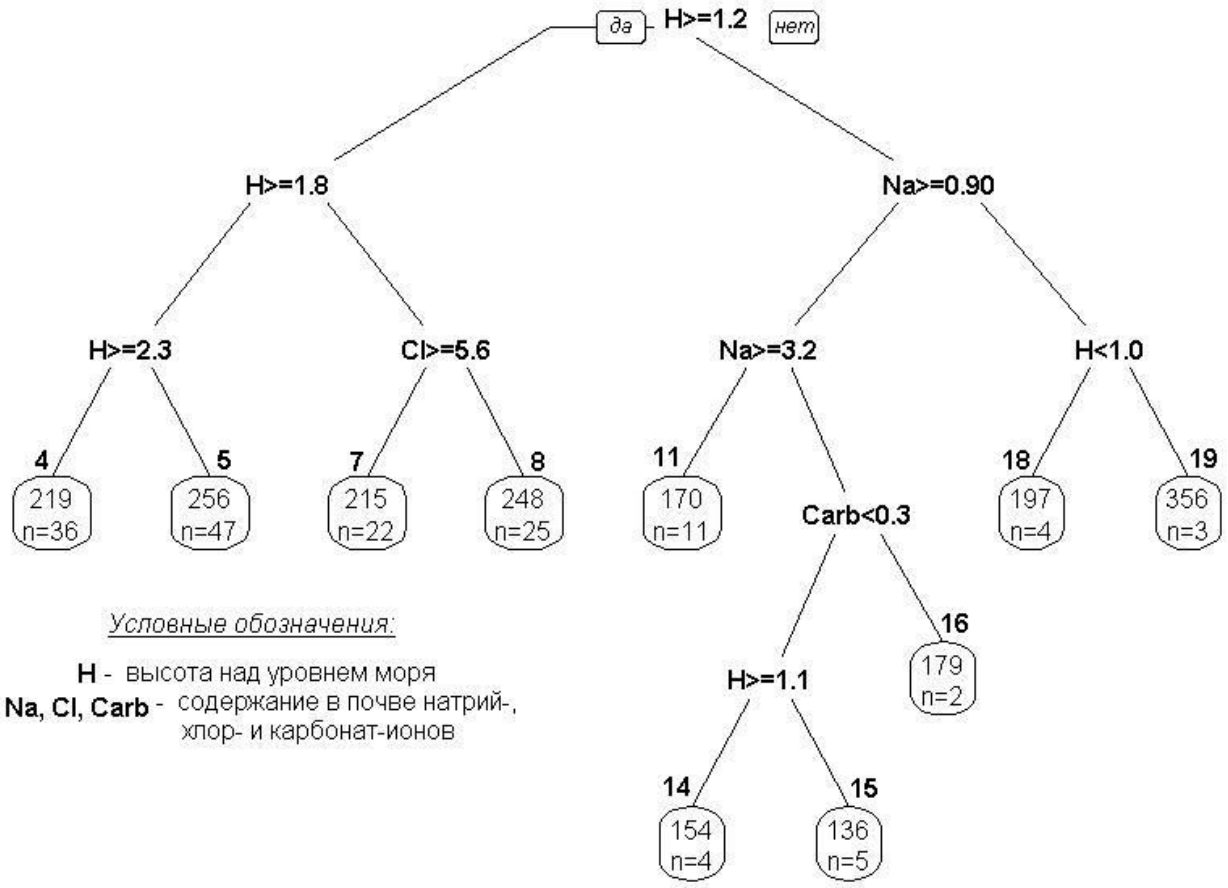


Рис. 6.15. Дерево MRT для классификации геоботанических описаний растительности в дельте р. Волга; в "листьях" указаны значения $\sqrt{SS_D}$ и число площадок в группе n

Механизм использования кросс-проверки для оптимизации деревьев MRT в целом аналогичен процедуре rpart, представленной в предыдущем разделе, и обеспечивает устойчивость модели в режиме предсказания. Для примера с обработкой геоботанических данных оптимальный размер дерева, оцениваемый по минимуму ошибки скользящего контроля CV_{er} равен 9 узлам разбиений, что приводит к образованию 10 групп (рис. 6.15).

Анализ многомерного отклика экосистемы с помощью деревьев MRT предоставляет исследователю много дополнительных возможностей интерпретации результатов. В первую очередь это связано с оценкой, какие виды и их ассоциации инициируют разбиение выборки на узлах дерева и предопределяют состав сформированных подмножеств объектов. На рис. 6.16 представлены диаграммы усредненных долей биомассы видов травянистых растений для трех характерных групп, обозначенных на рис. 6.15 "листьями" с номерами 4, 8 и 11.

Рис. 6.16. Доля отдельных видов травянистой растительности в трех группах площадок

Визуальный анализ диаграмм типа "резаного пирога" позволяет в простейших ситуациях легко установить, что, например, в группе 4 с пониженной влажностью появляются такие виды как шведка, солодка, петросимония.

В общем случае изучения многовидовых композиций задача выявления "индикаторных видов", которые являются экологическими указателями типов сообщества, условий окружающей среды или произошедших экологических изменений, решается расчетным путем. Для того чтобы установить, является ли произвольный вид индикаторным или фоновым, разработано два типа критериев ассоциативности (De Cáceres, Legendre, 2009): индикаторные индексы и коэффициенты корреляции. Индикаторный индекс d_{jk} (Dufrene , Legendre, 1997) вида j для k-й группы при делении исходной выборки на K кластеров является произведением относительной частоты f_{jk} и относительной средней популяционной плотности a_{jk} этого вида:

$$f_{jk} = \frac{\sum_{i \in k} p_{ij}}{n_k}; \quad a_{jk} = \frac{(\sum_{i \in k} y_{ij})/n_k}{\sum_{k=1}^{K}((\sum_{i \in k} y_{ij})/n_k)}; \quad d_{jk} = f_{jk} \times a_{jk},$$

где p_{ij} – событие (1/0) появления вида j на участке i; y_{ij} – численность или биомасса вида j на участке i; n_k – число измерений, попавших в k-ю группу.

Симметричный индекс индикаторной значимости вида равен $IndVal_j = \max[d_{jk}]$, т.е. он принимает максимальное значение (равное 100%), если экземпляры вида j встречаются во всех пробах только одной k-й группы. Естественно, что предположив закономерный характер связи встречаемости вида с одной из групп местообитаний, необходимо проверить нулевую гипотезу о случайности этой связи. Если выполнить пермутационный тест многократного перемешивания объектов в группах, то можно оценить p-значение $IndVal_j$ при справедливости H_0 как долю превышения его эмпирического значения над рандомизированными величинами, найденными при случайном распределении вида по участкам. Неоднократно описанная нами процедура бутстрепа позволит также оценить доверительные интервалы критерия. Как показано в табл. 6.5, индикаторная значимость индекса *IndVal* является статистически значимой лишь для 6 видов из 21.

Таблица 6.5. Виды травянистой растительности со статистически значимым индикаторным индексом *IndVal*

Вид	Группа с $\max[d_{jk}]$	Индекс *IndVal*	p-значение	Доверительный интервал
Скрытница	7	0.788	0.001	0.643 ÷ 0.918
Прибрежница	5	0.572	0.001	0.428 ÷ 0.711
Лебеда копьелист.	11	0.488	0.022	0.207 ÷ 0.736
Петросимония	4	0.417	0.021	0.25 ÷ 0.588
Клубнекамыш	11	0.38	0.033	0.19 ÷ 0.614
Рогоз	16	0.362	0.043	0 ÷ 0.991

К разделу 6.5:

```
load(file="Fito_Full.RData");  Species <- Species[,-8] ; Env <- Fito_Full[,27:36]
library(mvpart) ; library(rpart.plot) ; library(labdsv) ; library(indicspecies)
#  При выполнении функции mvpart выберите число групп щелчком мыши (например. 4)
spe.mvpart <- mvpart(data.matrix(Species) ~ ., Env, margin=0.08, cp=0,
       xv="pick", xval=nrow(Species), xvmult=100, which=4, bars = FALSE)
plot(spe.mvpart) ; text(spe.mvpart) ; summary(spe.mvpart) ; printcp(spe.mvpart1)
prp(spe.mvpart, extra=5)  #  Другой вариант отображения дерева
leaf.sum <- matrix(0, length(groups.mrt), ncol(Species))  # Доля биомассы видов в группах
(groups.mrt <- levels(as.factor(spe.mvpart$where))) ; colnames(leaf.sum) <- colnames(Species)
for(i in 1:length(groups.mrt)){   leaf.sum[i,] <-
              apply(Species[which(spe.mvpart$where==groups.mrt[i]),], 2, sum)} ; leaf.sum
#  Вывод диаграммы типа «разрезанный пирог»
par(mfrow=c(2,5)) ; for(i in 1:length(groups.mrt)){
       pie(which(leaf.sum[i,]>0), radius=1, main=c("Гр. № ", groups.mrt[i])) }
#  Расчет индикаторных значений IndVal
spe.MRT.indval <- indval(Species, spe.mvpart$where) ; spe.MRT.indval$pval    # P-значения
spe.MRT.indval$maxcls[which(spe.MRT.indval$pval<=0.05)]  # Номера групп с max[d]
spe.MRT.indval$indcls[which(spe.MRT.indval$pval<=0.05)] #  Значения IndVal для видов с p<0.05
#  Оценка доверительных интервалов IndVal
Ind.spe <- strassoc(Species, spe.mvpart$where, func="IndVal.g", nboot = 1000)
Ind.spe$stat^2 ; Ind.spe$lowerCI^2 ; Ind.spe$upperCI^2
```

6.6. Преобразование координат в геометрической морфометрии

Предметом многочисленных исследований в экологии, эволюции и систематике является выделение морфотипов изучаемой популяции и оценка их частотного распределения. При этом элементами морфопространства являются комбинации значений биометрических признаков для каждого наблюдаемого объекта (всей особи, части тела или отдельного органа), которые при их геометрической интерпретации соответствуют совокупностям точек на плоскости[12]. Рассмотрим в качестве примера внутрипопуляционную дифференциацию формы нижней челюсти мышевидных грызунов. На рис. 6.17а показано расположение ключевых точек, относительно которых выполнялись замеры. Если жевательную поверхность зубов расположить вдоль оси абсцисс, а передний край будет касаться оси ординат, то мы получим центр координат в метке 6. В качестве другого важного краниометрического элемента используется положение первого коренного зуба – метка 1.

При изучении морфометрической изменчивости широко используются подходы, основанные на традиционных многомерных методах статистики (Marcus at al., 1996; Claude, 2008). При этом в общем случае оценка различий между подмножествами объектов сводится к сравнению положения центроидов выделяемых групп и анализу компонентов дисперсий для всего множества координат меток. Необходимо найти ответы на два вопроса: (а) являются ли различия между группами статистически значимыми, что является целью многомерного дисперсионного анализа MANOVA, подробно рассмотренного в разделе 4.7, и (б) на основе каких решающих правил или моделей с использованием информативных переменных можно идентифицировать принадлежность неизвестного объекта к той или иной группе.

В табл. 6.6 приведен статистический анализ MANOVA значимости различий между группами красной полевки (всего 837 особей) на основе формы нижней челюсти, представленной на рис. 6.17*а* декартовыми морфометрическими координатами меток. Группировка выполнялась по двум категориям: "Пол" (419 самок и 418 самцов) и "Местообитание" (популяция в пределах 5 километровой зоны Байкальского ЦБК и за ее пределами), влияние которых рассматривалось как отдельно, так и в рамках двухфакторного анализа. Для всех вариантов расчета и использованных критериев была

[12] Или, в общем случае, в 3D-пространстве или гиперобъеме

установлена статистически значимая зависимость изменчивости среднегрупповых координат точек от обоих факторов.

Таблица 6.6. Многомерный анализ MANOVA кранометрических промеров нижней челюсти красной полевки в декартовых координатах в зависимости от пола и местообитания (в числителе – результаты однофакторного, в знаменателе – двухфакторного анализа)

Тестовая статистика	Фактор "Пол"			Фактор "Местообитание"		
	Критерий	*F*-аппрокс.	*p*-значение	Критерий	*F*-аппрокс.	*p*-значение
Пиллая (Pillai's Trace)	0.048 0.048		0.00414 0.00407	0.029 0.028		0.00645 0.00878
Λ Уилкса (Wilks' Lambda)	0.952 0.952	2.042 2.047	0.00418 0.00411	0.971 0.972	2.471 2.38	0.00645 0.00878
Хотеллинга (Hotelling-Lawley)	0.0496 0.0497	2.043 2.045	0.00423 0.00415	0.0299 0.0289	2.471 2.381	0.00645 0.00878
Корень Роя (Roy's Root)	0.0291 0.0292	2.408 2.405	0.008 0.0081	0.0299 0.0289	2.471 2.38	0.00645 0.00878

Однако использование в морфометрических исследованиях исходных данных, представленных в виде декартовых координат точек, корректно лишь в том случае, когда различия в размере объектов не вносят дополнительной систематической ошибки, либо сама размерная вариация особей является предметом изучения. Если смысл задачи сводится к анализу изменчивости формы объектов, то необходимо провести изометрические преобразования исходной системы координат, благодаря которым из дальнейшего анализа исключается "размерный фактор". К таким преобразованиям относятся смещение (translation) и поворот (rotation) системы координат, при которых происходит сдвиг всей структуры, а также масштабирование (scaling), при котором пропорционально меняются расстояния между метками. Геометрическая форма объектов при изометрических преобразованиях не меняется. Набор специфических алгебраических техник, позволяющих создать пространство структур, инвариантное к размерам, в сочетании с многомерным анализом координат меток составляют принципы и методы *геометрической морфометрии* (Dryden, Mardia, 1998; Павлинов, Микешина, 2002).

Частным случаем выравнивания размеров форм при сохранении исходных относительных расстояний между метками является использование *букштейновых* координат (Bookstein coordinates – Bookstein, 1991). Для их определения из множества меток выбираются две, которые определяют так называемую базовую линию (baseline): в зависимости от задачи исследования это могут быть метки, наиболее удаленные друг от друга или, например, обозначающие какую-либо функционально значимую структуру. Выбранным меткам присваиваются координаты (-0.5, 0) и (0.5, 0), а прочие метки переопределяются как вершины треугольников с общим основанием, заданным базовой линией; соответственно пересчитываются их координаты. На рис. 6.17*б*, *в* представлен итог процедуры наложения (superimposition) меток в букштейновых координатах для всех 837 особей, а также средние формы нижней челюсти полевки красной для самцов и самок, которые, к нашему огорчению, оказались визуально неразличимыми.

Результат преобразования Букштейна в значительной мере зависит от удачного выбора базовой линии, относительно которой, собственно, и осуществляется масштабирование и поворот. Если принять в качестве базовой, например, линию между метками 2 и 6, то конфигурация формы окажется сильно деформированной. Этого недостатка лишено *кендэллово* пространство форм (Kendall's shape space), метрика которого определяется прокрустовой дистанцией. Это позволяет задать полный набор геометрических характеристик для анализа отношений между формами, поэтому большинство многомерных методов геометрической морфометрии основано на статистических оценках расстояний и направлений в кендэлловом пространстве.

На основе *прокрустовой оптимизации* выполняются операции смещения, масштабирования и вращение всех совмещаемых форм таким образом, чтобы сумма квадратов расстояний между соответствующими выровненными метками была бы минимальной. При этом возможны различные варианты итеративного алгоритма подгонки: например, в пакете shapes среды R предложено три версии прокрустова анализа – простого, обобщенного и взвешенного (Ordinary, Generalised, Weighted Procrustes fit, подробности см. Goodall, 1991; Dryden, Mardia, 1998).

Вариация форм после преобразования и наложения оценивается как величина *прокрустовых остатков*, которые рассматриваются как специфические переменные, используемые непосредственно в многомерной статистике. Эти остатки (или прокрустовы координаты) представляют собой совокупность векторов расстояний между образами каждого экземпляра, максимально совмещенными с "эталоном" путем вышеперечисленных трансформаций, и соответствующими метками эталонной конфигурации – рис. 6.17*г*.

После прокрустовых преобразований можно получить результаты сопоставления усредненных форм для групп объектов, в которых влияние размерного фактора исключено. Некоторой специальной модификацией критерия Хотеллинга T^2, используемого для сравнения центроидов двух многомерных выборок и описанного в разделе 4.7, являются F-тест Гудолла (Goodall's F-test – см. Dryden, Mardia, 1998) и статистика T^2 Джеймса (James T^2 statistic – см. Amaral et al., 2007), которые учитывают анизотропность системы морфометрических координат и оценивают эквивалентность средних форм по группам на основе прокрустовых расстояний.

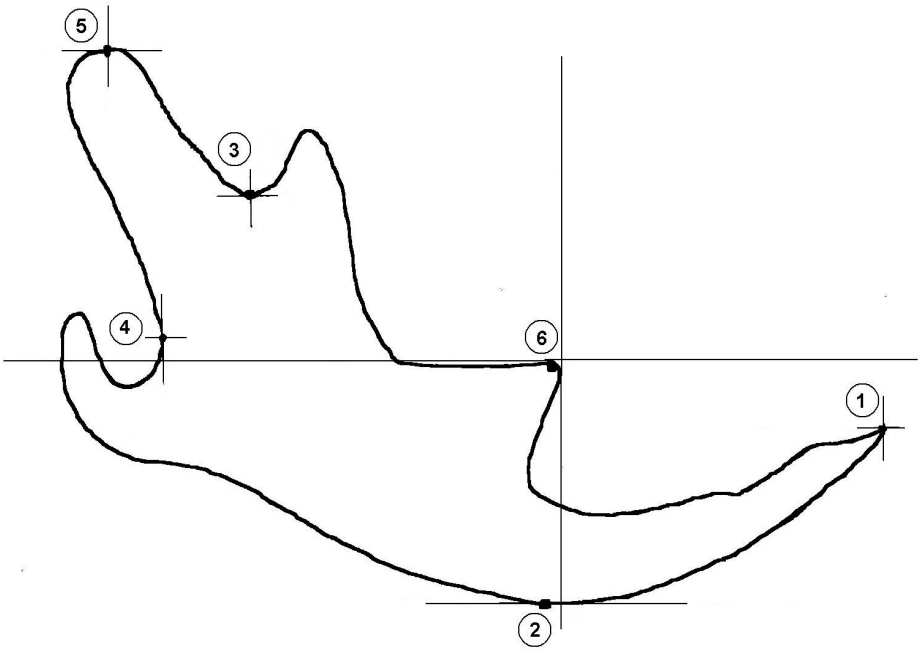
С использованием функции testmeanshapes(…) может быть выполнена проверка нулевой гипотезы об эквивалентности двух средних форм с использованием статистик Хотеллинга, Гудолла и Джеймса – см. табл. 6.7. Оценка p-значения может быть выполнена как на основе аппроксимации F-распределением, так и с использованием рандомизационной процедуры (например, после 1000 итераций случайного перемешивания строк данных между двумя группами).

Таблица 6.7. Статистический анализ половых и популяционных различий формы нижней челюсти красной полевки после прокрустова преобразования координат меток;
p-парам – оценка значимости на основе аппроксимации F-распределения,
p-ранд – с использованием пермутационного теста

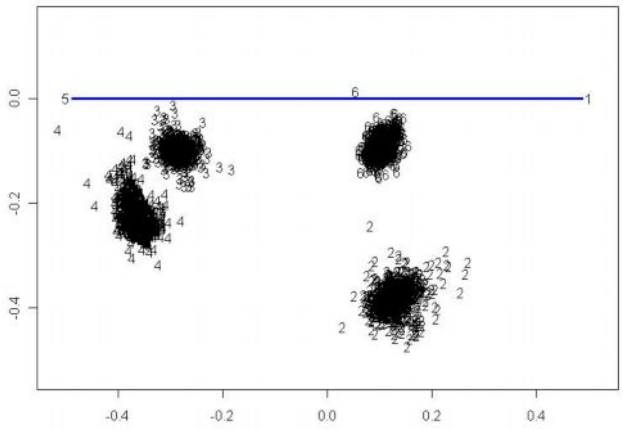
Название статистики	Фактор "Пол"			Фактор "Местообитание"		
	Статистика	p-парам	p-ранд	Статистика	p-парам	p-ранд
Хотеллинга T^2	2.26	0.0214	0.0396	0.752	0.644	0.657
Гудолла G	1.922	0.0523	0.0594	0.614	0.766	0.698
Джеймса T^2	18.2	0.021	0.0396	5.93	0.661	0.667

Результаты тестирования показывают, что нет оснований отклонить H_0 об идентичности формы нижней челюсти полевки на разном удалении от БЦБК, т.е. различия, полученные MANOVA в табл. 6.6, объясняются лишь разными размерными характеристиками черепа грызунов. Вывод о влиянии гендерного фактора на форму объекта с целом подтвердился, но его статистическая значимость оказалась на пороге общепринятого уровня доверительности $p = 0.05$. Для специалистов может оказаться полезным величина римановского расстояния (Riemenian distance) между формами челюстей самок и самцов, равное 0.00337, что свидетельствует о низкой инвариантности отражения.

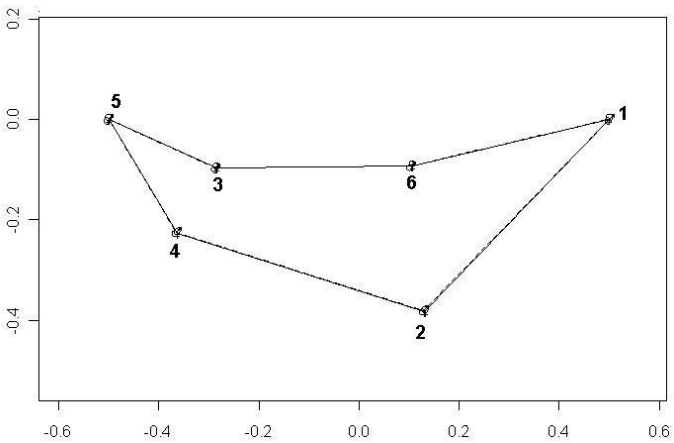
Эффективным методом выявления основных тенденций изменчивости форм является анализ главных компонент PCA. На рис. 6.17*д* показаны направления деформаций меток нижней челюсти грызунов в пространстве первой главной компоненты, которая объясняет 32.7% общей вариации. Кроме того PCA является мерой преодоления "проклятия размерности" в многомерном анализе при большом числе координат формы.



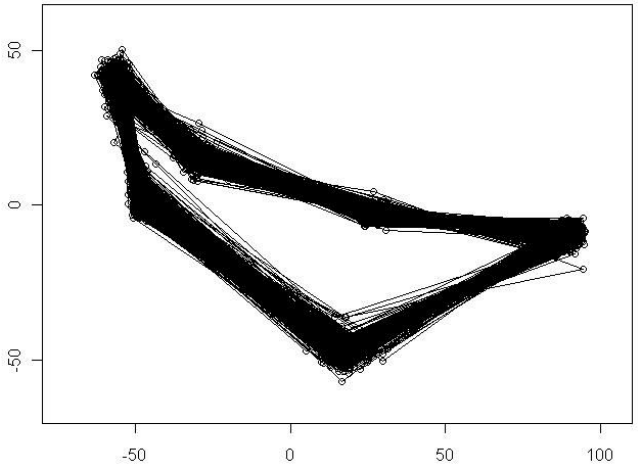
а) Расположение стандартных точек на нижней челюсти полевки красной



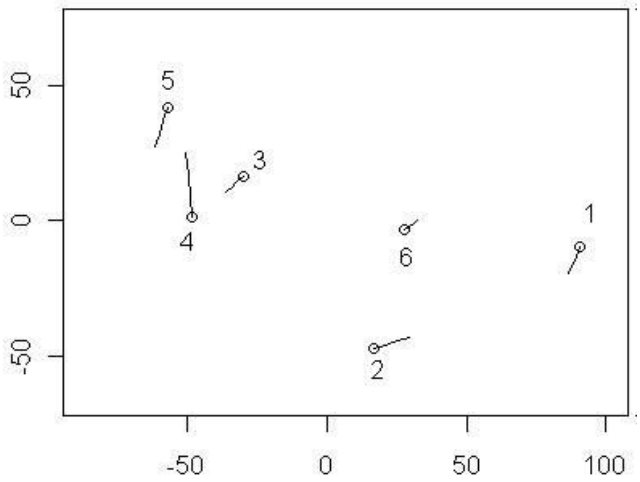
б) базовая линия и вариация букштейновых координаты



в) средние формы челюсти самок и самцов в букштейновых координатах



г) вариация формы челюсти в прокрустовых координатах



д) направление и выраженность сдвига меток для первой главной компоненты

Рис. 6.17. Использование методов геометрической морфометрии для анализа мандибул грызунов

Как вариант, при выполнении многомерного дисперсионного анализа MANOVA в качестве отклика можно использовать матрицу счетов (PCA scores) первых 8 главных компонент, объясняющих 99.8% вариации прокрустовых остатков, что позволит избежать коллинеарности векторов при формировании ковариационной матрицы. В этих условиях при формировании групп по полу грызунов гендерные отличия мандибул оказалась достаточно велики, чтобы отклонить H_0 (доля внутригрупповой вариации переменных форм $\Lambda = 0.964$, $F = 1.92$ при использовании статистики Хотеллинга, $p = 0.015$). Отличия в форме челюсти между самками и самцами оказались значимыми и при использовании статистики Пиллая с тем же *p*-значением. С другой стороны, доля объясненной суммы квадратов в полной вариации отклика, рекомендованная для биологических исследований Н.А.Плохинским, оказалась удручающе мала $\eta^2 = 0.0183$.

К разделу 6.6:

```
library(xlsReadWrite) ; GM <- read.xls("Ruts.xls", sheet = 2, rowNames=TRUE)
#  Функция перегруппировки исходной таблицы в формат конфигурационной матрицы m×2×n
conf_mat <- function (df, m, n)  { cm <- array(rep(0,m*2*n),dim=c(m,2,n))
for (i in 1:n) { for (j in 1:m) {cm[j,1,i] = df[i,j*2-1]; cm[j,2,i] = df[i,j*2]}} ; cm }
#  Формирование трех матриц: самок и самцов отдельно и по всей выборке
m <- 6 ; ruts.cm <- conf_mat(GM[,3:14], m, nrow(GM))
A.cm <- ruts.cm[,,which(GM$sex =="F")] ; B.cm <- ruts.cm[,,which(GM$sex =="M")]
#  Для группировки по местообитаниям снимите ниже знак комментария и повторите все расчеты
#  A.cm <- ruts.cm[,,which(GM$region ==1)] ; B.cm <- ruts.cm[,,which(GM$region != 1)]
dim(A.cm) ; dim(B.cm) ; dim(ruts.cm)  ; library(shapes)
#  MANOVA  по меткам в декартовой системе координат
model11<-manova(cbind(X1,Y1,X2,Y2,X3,Y3,X4,Y4,X5,Y5)~sex, data=GM)
model12<-manova(cbind(X1,Y1,X2,Y2,X3,Y3,X4,Y4,X5,Y5)~region, data=GM)
model<-manova(cbind(X1,Y1,X2,Y2,X3,Y3,X4,Y4,X5,Y5)~sex*region, data=GM)
anova(model) ; summary(model,test="W"); summary(model,test="H") ; summary(model,test="R")
#  Расчет координат Букштейна и и вывод графиков 'б' и 'в' на рис. 6.17
book.ruts <- bookstein2d(ruts.cm, 5, 1)
segments(-0.5,0,0.5,0, col="blue",lwd=3) ; book.ruts$mshape
book.A <-bookstein2d(A.cm, 5, 1) ; book.B <-bookstein2d(B.cm, 5, 1)
book.A$mshape ; book.B$mshape
plotshapes(book.A$mshape,joinline=c(1,2,4,5,3,6,1), symbol="")
lines(book.B$mshape[c(1,2,4,5,3,6,1),],lty=2)
text(book.A$mshape,  "\\VE", family = "HersheySerif") #  Метки со знаком  Female
text(book.B$mshape,  "\\MA", family = "HersheySerif")  #  Метки со знаком  Male
#  Обобщенный прокрустов анализ
proc.ruts =procGPA(ruts.cm) ; proc.A <-procGPA(A.cm)  ;  proc.B <-procGPA(B.cm)
plotshapes(proc.ruts$rotated, joinline=c(1,2,4,5,3,6,1))  #  рис. 6.17г
plotshapes(proc.ruts$mshape, joinline=c(1,2,4,5,3,6,1)) # средняя (эталонная) форма
mean(proc.ruts$size) ; mean(proc.A$size) ; mean(proc.B$size) #  размеры центроида
riemdist(proc.A$mshape, proc.B$mshape)  #  Римановское расстояние
testmeanshapes(A.cm, B.cm, resamples=1000)  #  Тест на значимость отличий форм по группам
shape_variables<-t(matrix(proc.ruts$rotated,m*2,nrow(GM)))
#  Вывод диаграмм PCA в разных вариантах
shapepca(proc.ruts, pcno = c(1, 2), type = "v",mag=3)
shapepca(proc.ruts, pcno = c(1, 2), type = "r",mag=3)
#  Выполняем PCA-анализ  с помощью функции rda() пакета  vegan
library(vegan) ; pca.ruts<-rda(shape_variables) ; summary(pca.ruts)
plot(proc.ruts$scores[,1],pca.ruts$CA$u.eig[,1]) #  Убеждаемся в идентичности значений счетов
# и рисум диаграмму распределения особей в координатах двух главных компонент
plot(pca.ruts$CA$u.eig[,1],pca.ruts$CA$u.eig[,2],pch=16, col=c("red","blue")[GM$sex])
# Для анализа MANOVA прокрустовых координат используем 8 главных компонент
response_matrix<-cbind(pca.ruts$CA$u.eig[,1:8]) ; model<-manova(response_matrix~GM$sex)
anova(model) ; summary(model,test="W"); summary(model,test="H") ; summary(model,test="R")
library(heplots) ; etasq(model)  # Вычисляем корреляционное отношение η²
save(file="GM.RData",GM, shape_variables, response_matrix) #  Вывод в файл  для раздела 6.7
```

6.7. Дискриминантный анализ, логистическая регрессия и метод опорных векторов

Рассмотрим несколько алгоритмов распознавания, основанных на использовании принципа разделения (*R*-модели). Они различаются главным образом заданным классом поверхностей, среди которых выделяются такие, которые наилучшим образом разделяют объекты разных классов. В качестве примера продолжим рассмотрение задачи геометрической морфометрии и будем изыскивать возможность определения пола грызунов по форме нижней челюсти (вряд ли эта проблема имеет практическое применение и приводится здесь чисто в методологическом плане). В качестве независимых переменных модели используем 8 главных компонент $PC1 \div PC8$, полученных при анализе PCA прокрустовых координат и определяющих основные компоненты вариации формы, а также число перфораций $F1 \div F3$ костей черепа перед верхней челюстью, в глазничной впадине и под нижней челюстью соответственно.

Дискриминантный анализ является разделом многомерной статистики, который позволяет оценить различия между двумя и более группами объектов, описанных одновременно множеством **X** наблюдаемых переменных. Дискриминантный анализ (или анализ дискриминантных функций) – общий термин, относящийся к нескольким тесно связанным статистическим процедурам.

В дискриминантном анализе обычно принимается основное предположение, что описания объектов каждого *k*-го класса представляют собой реализации многомерной случайной величины, распределенной по нормальному закону $N_p(\boldsymbol{\mu}_k; \boldsymbol{\Sigma}_k)$ со средним $\boldsymbol{\mu}_k$ и дисперсией $\boldsymbol{\Sigma}_k$ (индекс *p* указывает на размерность признакового пространства). С учетом этого, на основании обучающей выборки формируется решающее правило, позволяющее отнести новый объект, представленный также в *p*-мерном пространстве, к классу *k*, у которого есть самая высокая плотность вероятности ("уровень правдоподобия") в этой точке.

Линейный анализ дискриминантных функций (LDA) в качестве решающего правила вычисляет координаты $(k - 1)$ многомерных плоскостей

$$Z_k(\boldsymbol{x}) = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p ,$$

используемых для разделения ("дискриминации") классов. Анализ ведется по следующим формулам:

- находятся ковариационные матрицы $\mathbf{C}_k$ объектов каждого *k*-го класса из *g* и проводится их объединение в расчетную ковариационную матрицу **C**;

$$\mathbf{C}_k = \frac{1}{n_k - 1}\sum_{i=1}^{n_k}(\boldsymbol{x}_{ik} - \boldsymbol{\mu}_k)^{\mathrm{T}}(\boldsymbol{x}_{ik} - \boldsymbol{\mu}_k); \quad \mathbf{C} = \frac{1}{N - g}\sum_{k=1}^{g}(n_k - 1)\mathbf{C}_k; \qquad k = 1, 2, \ldots g;$$

- по формуле $\boldsymbol{\beta} = \mathbf{C}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$ вычисляется вектор коэффициентов $\{\beta_1, \ldots, \beta_p\}$ уравнения разделяющей гиперплоскости между классами 1 и 2;
- обобщенное расстояние Махаланобиса или дистанция в многомерном пространстве признаков между центроидами двух групп объектов оценивается как $D^2 = \boldsymbol{\beta}^{\mathrm{T}}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2)$.

Интерпретацию LDA для случая $p = 2$, когда гиперплоскость сводится к линии LD_2 см. на рис. 6.1б.

Таким образом, в LDA кроме предположения о нормальности распределения данных в каждом классе, выдвигается еще и более серьезное предположение о статистическом равенстве внутригрупповых матриц дисперсий и корреляций. При искусственном объявлении ковариационных матриц $\mathbf{C}_k$ статистически неразличимыми могут оказаться отброшенными наиболее важные индивидуальные черты, имеющие большое значение для хорошей дискриминации.

Для рассматриваемого примера при использовании полного комплекта переменных уравнение разделяющей плоскости будет иметь вид:

$$Z_2(\boldsymbol{x}) = -0.247F1 - 0.091F2 - 0.107F3 + 0.429PC1 + 1.85PC2 + 3.15PC3 - 2.6PC4 - \\ - 9.47PC5 + 0.223PC6 - 4.75PC7 - 15PC8,$$

а расстояние Махалонобиса между центроидами двух групп объектов $D^2 = 0.442$. Если после подстановки в это уравнение значений морфометрических переменных конкретного животного значение величины Z окажется больше 0, то оно определяется как самец, если меньше 0 – как самка. Доля ошибок модели при опознании примеров исходной выборки представлена в табл. 6.8 и будет обсуждаться позднее.

Важной характеристикой прогнозирующей эффективности модели является ошибка при кросс-проверке. В функции lda(…) пакета MASS заложена реализация скользящего контроля (leave-one-out CV), т.е. из исходной выборки поочередно отбрасывается по одному объекту, строится n моделей дискриминации по $(n - 1)$ выборочным значениям, а исключенная реализация каждый раз используется для классификации. Доли ошибочных опознаний при такой процедуре также приведены в табл. 6.8.

Наконец, было бы естественным задаться вопросом, какие из имеющихся 11 признаков являются информативными при разделении, а какие – сопутствующим балластом. Шаговая процедура выбора переменных при классификации, реализованная функцией stepclass(…) пакета klaR, основана на вычислении сразу четырех параметров качества модели-претендента: а) индекса ошибок (correctness rate), б) точности (accuracy), основанной на евклидовых расстояниях между векторами "факта" и "прогноза", в) способности к разделимости (ability to seperate), также основанной на расстояниях, и г) доверительных интервалах центроидов классов. При этом все эти параметры оцениваются в режиме многократной кросс-проверки.

Проведенная селекция статистически значимых переменных для рассматриваемого примера привела к весьма лаконичной модели:

$$Z_2(\boldsymbol{x}) = 11.18PC5 - 17.6PC8$$

Квадратичный дискриминантный анализ (QDA) является нелинейным обобщением метода LDA при разделении данных на два или более классов. В качестве решающего правила используется квадратическая функция, которая вычисляется на основе внутригрупповых ковариационных матриц каждого класса:

$$Z_k(\boldsymbol{x}) = -0.5(\mathbf{x} - \boldsymbol{\mu}_k)^T \mathbf{C}_k\text{-}1(\boldsymbol{x} - \boldsymbol{\mu}_k) - 0.5\ln|\mathbf{C}_k| + \ln P(c_k),$$

где $|\mathbf{C}_k|$ - детерминант ковариационной матрицы, $P(c_k)$ – вероятность появления объектов k-го класса. Экзаменуемый объект также относится к тому классу, которому соответствует максимальное значение Z_k.

Квадратичный дискриминантный анализ весьма эффективен, когда разделяющая поверхность между классами имеет ярко выраженный нелинейный характер (например, параболоид или эллипсоид в 3D-случае). Однако он сохраняет большинство недостатков LDA: использует предположение о нормальности распределения и не работает, когда матрицы ковариаций вырождены, например, при большом числе переменных. Другим недостатком QDA является то, что уравнение разделяющей гиперповерхности выражено в неявном виде и не может быть использовано для "объяснения".

Логистическая регрессия (LR) – наиболее общий подход к моделированию значений отклика, заданного альтернативной шкалой (0/1). Сущность этого метода на примере функции одной переменной рассматривалась нами ранее в разделе 3.6.

Если $\boldsymbol{x}$ – вектор предикторных значений, состоящий из p независимых переменных, то вероятность P того, что отклик y примет значение 1, может быть описана моделью

$$p(\boldsymbol{x}) \equiv P(y = 1|\boldsymbol{x}),$$

где $p(\boldsymbol{x})$ – линейная функция $\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p$,

которая после оценки коэффициентов регрессии β_i будет возвращать искомую вероятность на интервале [0,1]. Для того, чтобы обеспечить робастность процедуры подбора коэффициентов, модель переписывают как функцию *логита*

$$g(x) = \log\left(\frac{p(x)}{1 - p(x)}\right) = \beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p,$$

которая линейна относительно предикторов x_i. Однако, независимо от этого, результат $\hat{\beta}_0 + \hat{\beta}_1 x_1 + ... + \hat{\beta}_p x_p$ трансформируется обратно в вероятность $p(\boldsymbol{x})$ между 0 и 1.

Неизвестные параметры модели (т.е. коэффициенты $\beta_0, \beta_1, ..., \beta_p$) обычно находятся с использованием функции максимума правдоподобия

$$\prod_{i=1}^{n} \left\{ p(\bar{x})^{\,y_i} \left[1 - p(\bar{x}_i) \right]^{1-y_i} \right\},$$

для которой справедливо выражение $P(Y_1 = y_1, ..., Y_n = y_n \mid \vec{x}_1, ..., \vec{x}_n)$.

Для рассматриваемого примера при использовании полного комплекта переменных уравнение логита будет иметь вид:

$g(\boldsymbol{x})$ = 0.295 - 0.076*F1* - 0.027*F2* - 0.033*F1* + 0.131*PC1* + 0.566*PC2* + 0.983*PC3* -
- 0.802*PC4* - **2.91*PC5*** + 0.05*PC6* - 1.45*PC7* - **4.63*PC8*** (критерий AIC = 1165),

где жирным шрифтом отмечены статистически значимые термы с $p < 0.05$. Нетрудно заметить, что это уравнение с точностью до масштабирующего множителя очень похоже на приведенное выше уравнение разделяющей гиперплоскости LDA.

Селекцию информативных переменных модели можно осуществить с использованием стандартной процедуры step(…), в результате чего получаем компактное уравнение логита:

$$g(\boldsymbol{x}) = -0.00265 - 2.95PC5 + 4.67PC8,$$

доставляющее минимум критерию Акаике AIC = 1152. Подробности статистического анализа и идентификации обобщенных линейных моделей см. в разделах 3.6, 4.1 и 4.6.

Метод *опорных векторов* или SVM (Support Vector Machine), называемый также алгоритмом "обобщенного портрета", разработан в серии работ В.Н.Вапника (Вапник, Червоненкис, 1974; Алгоритмы и программы…, 1984). Это метод, используемый как для классификации, так и для регрессии, заключается в решении задачи минимизации эмпирических кусочно-линейных функций штрафа. Опыт применения SVM для оценки класса качества вод малых рек по макрозообентосу подробно описан нами в монографии (Шитиков и др., 2005).

В простейшем случае линейной дискриминации точек двух классов в векторном пространстве обучающей выборки находятся параметры w_i и b уравнения разделяющей поверхности $z_k(x) = \sum_{i=1}^{p} w_i x_i + b$, которая максимально удалена от двух специально подобранных подмножеств опорных векторов – см. рис. 6.18. Опорные векторы являются критическими элементами процесса обучения, а все остальные точки являются "резервуаром" для оптимизации их числа. При отсутствии линейной разделяемости решается задача квадратичного программирования, оптимизирующая соотношение между шириной полосы разделения и суммой штрафов за исключение опорных векторов, препятствующих правильному распознаванию.

Если линейный классификатор показывает неудовлетворительные результаты, то разделяющую поверхность $\bar{w}^T \varphi(\bar{x}) = -b$ можно найти в расширенном пространстве признаков $\varphi(\bar{x})$, полученном в результате нелинейного преобразования с использованием полинома, радиальной базисной функции или сигмоида. Разумеется, в случае нелинейной модели возникает дополнительная задача, как найти наилучшую функцию ядра (например, наиболее подходящую степень полинома оптимальной разделяющей поверхности).

Использование линейной модели SVM позволяет вывести для каждой пары исходных переменных двухмерную диаграмму (plot), показывающую положение следа разделяющей гиперплоскости и ошибки распознавания. Рис. 6.19 иллюстрирует, насколько непростая задача была поставлена перед алгоритмами классификации, чтобы правильно распознать пол грызуна по форме нижней челюсти.

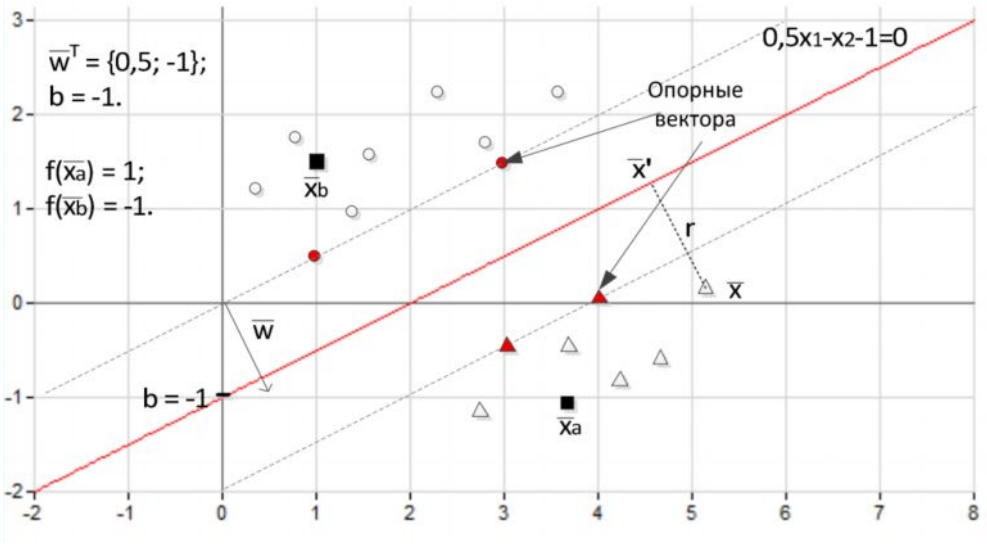


Рис. 6.18. Иллюстрация метода "обобщенного портрета". Опорные векторы определяют максимальную ширину разделяющей полосы. $\bar{w}^T$ – вектор нормали, задающей уравнение разделяющей поверхности $\bar{w}^T\bar{x}=-b$, расстояние от которой до произвольной точки равно *r*. Экзаменуемая точка $\bar{x}_a$ имеет значение классификатора $f(\bar{x}_a) = sign(\bar{w}^T\bar{x}_a + b) = +1$ и относится к классу "треугольников". Другая точка $\bar{x}_b$, для которой $f(\bar{x}_b) = -1$, опознается как "кружочек".

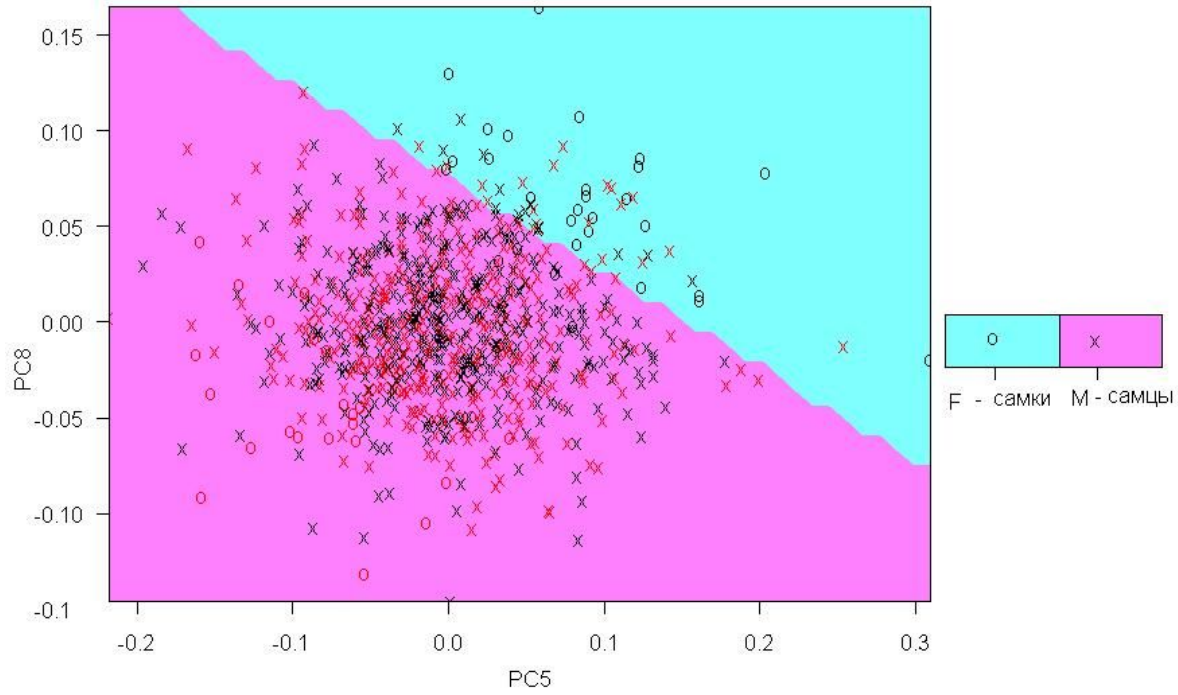


Рис. 6.19. Положение разделяющей гиперплоскости, полученной методом опорных векторов, в пространстве главных компонент PC5-PC8 переменных формы нижней челюсти грызунов; красным цветом отмечены ошибки распознавания.

При использовании метода опорных векторов построение разделяющей гиперплоскости для нашего примера осуществлялась по 782 точкам, т.е. 55 векторов исходной выборки были исключены как "злостно" препятствующие классификации, благодаря чему модель распознавания получила дополнительную устойчивость, в том числе, при скользящем контроле. Сокращение признакового пространства с 11 до 3 переменных (F1, PC5, PC8) также не сказалось на качестве распознавания. Авторы

рекомендуют обратить внимание на широкое распространение и постоянное развитие алгоритмов SVM, современные версии реализации которых представлены в пакетах klaR и penalizedSVM среды R.

Для решения предъявленного примера все использованные модели распознавания LDA, QDA, LR и SVM показали весьма высокую вероятность ошибочного прогноза (табл. 6.8), что, вероятно, можно расценивать как неудачу с предметной точки зрения. Несмотря на это, наша задача - обратить внимание на следующие общие методические закономерности, важные при выборе алгоритмов классификации и анализе результатов их работы:

- ошибка кросс-проверки E_{CV} всегда превышает внутреннюю ошибку модели E_S на самой обучающей выборке и является объективной характеристикой качества распознавания на внешнем дополнении;
- сокращение размерности модели за счет выбора комплекса информативных переменных может увеличить ошибку E_S, но, как правило, снижает ошибку скользящего контроля E_{CV};
- использование "продвинутых" нелинейных поверхностей разделения (QDA в табл. 6.8) часто приводит к формированию "переопределенных" моделей, эффективных на самой обучающей выборке, но не столь успешных при экзамене неизвестных объектов.

Таблица 6.8. Оценка прогнозирующей эффективности различных методов распознавания пола грызунов по форме нижней челюсти; *Err-S* – доля ошибочной классификации на обучающей выборке; *Err-CV* – то же, при скользящем контроле

Использованный алгоритм	Полная модель		С селекцией переменных	
	Err-S	*Err-CV*	*Err-S*	*Err-CV*
Линейный дискриминантный анализ LDA	0.442	0.473	0.452	0.452
Квадратичный дискриминантный анализ QDA	0.388	0.473	0.453	0.466
Логистическая линейная регрессия LR	0.44	0.473	0.452	0.454
Метод опорных векторов SVM	0.432	0.432	0.433	0.434

К разделу 6.7:

```
load(file="GM.RData")   # Используем данные, сохраненные в скрипте к разделу 6.6
Gr <- as.factor(GM[,1]) ; DS <- cbind(GM[c(1,15:17)],response_matrix) ; attach(DS)
#  ------------- Дискриминантный  анализ  LDA  и QDA
# Функция вывода результатов классифицирования
Out_CTab <- function (model, group, type="lda") {
       # Таблица сопряженности  "Факт/Прогноз" для модели по всей обучающей выборке
       classified<-predict(model)$class  ; t1 <-  table(group,classified)
       # Ошибка классифицирования  и расстояние Махалонобиса
       Err_S <- mean(group != classified) ; mahDist <- NA
       if (type=="lda") { mahDist <- dist(model$means %*% model$scaling) }
       # Таблица  "Факт/Прогноз" и ошибка при тестировании модели по скользящему контролю
       t2 <-  table(group, update(model, CV=T)$class->LDA.cv)
       Err_CV <- mean(group != LDA.cv) ; Err_S.MahD  <-c(Err_S, mahDist)
       Err_CV.N <-c(Err_CV, length(group)) ; cbind(t1, Err_S.MahD, t2, Err_CV.N)}
# --- Выполнение расчетов
lda.all <- lda(sex ~ ., DS); Out_CTab(lda.all, Gr)
qda.all <- qda(sex ~ ., DS) ; Out_CTab(qda.all, Gr, type="qda")
# Для селекции переменных используем функцию stepclass() пакета  klaR
library(klaR)  ; stepclass(sex ~ ., data=DS,  start.vars = "PC5", method="lda")
lda.step <- lda(sex ~ PC5 + PC8, DS) ; Out_CTab(lda.step, Gr)
stepclass(sex ~ ., data=DS,  method="qda")
qda.step <- qda(sex ~ PC4, DS) ; Out_CTab(qda.step, Gr, type="qda")
#   ------------- Логистическая линейная регрессия с биномиальной связью
# Функция вывода результатов классифицирования
```

```
Out_LR <- function (model, group) {
       mp <- predict(model, type="response") ; posterior <- data.frame(F=1-mp, M=mp)
       classified <- colnames(posterior)[apply(posterior,1,which.max)]
       CTab <- table(Факт=group, Прогноз=classified) ; Err_S <- mean(group!= classified)
       return(list(CTab=CTab, Err_S=Err_S))}
summary(glm(sex ~ ., DS, family=binomial) -> lr.all) ; Out_LR(lr.all, Gr)
library(boot)  #  Для оценки ошибки кросс-проверки используем функцию cv.glm()
cost <- function(r, pi = 0) mean(abs(r-pi) > 0.5)
(Err_CV <- cv.glm(DS, lr.all, cost)$delta)
# Для селекции переменных используем функцию step ()
lr.step <- step(lr.all) ; Out_LR(lr.step, Gr) ; cv.glm(DS, lr.step, cost)$delta
#   ------------- Метод опорных векторов SVM
#   Функция вычисления ошибки кросс-проверки
CVsvm <- function(x, y) {
     n <- nrow(x) ; Err_S  <- 0 ; for(i in 1:n)
       {svm.temp <- svm(x=x[-1,], y = y[-1],  kernel = "linear")
       if (predict(svm.temp, newdata=x[i,]) != y[i]) Err_S <- Err_S +1 }  ;  Err_S/n }
svm.all <-  svm(formula = sex ~ ., data = DS,  kernel = "linear") ; CVsvm(DS[,2:12],Gr)
# А здесь кросс-проверка (cross=10) используется только для подбора параметров модели
svm.cv <- svm(formula = sex ~ ., data = DS, cross=10, kernel = "linear")
table(Факт=Gr, Прогноз=predict(svm.cv))  ; mean(predict(svm.cv) != Gr )
stepclass(sex ~ ., data=DS,  method="svmlight")  # Для селекции  используем другую версию SVM
svm.fit <-  svm(formula = sex ~ F1+PC5+PC8, data = DS,  kernel = "linear")
table(Факт=Gr, Прогноз=predict(svm.fit))
mean( predict(svm.fit) != Gr )  ;  CVsvm(DS[,c(2,9,12)],Gr)
```

6.8. Метод *k* ближайших соседей и использование нейронных сетей

В общем случае описанные выше методы могут быть использованы для построения решающих правил распознавания k классов, где $k > 2$.

В качестве примера рассмотрим задачу географического районирования популяций гадюки обыкновенной, включающей два подвида *V. b. berus* и *V. b. Nikolskii*. Выборку из отловленных животных (100 экземпляров) разделим на три группы: "северную" **N** (Пермский край и Вологодская обл.), "западную" **W** (Липецкая и Пензенская обл.) и "южную" **S** (Самарская и Саратовская обл.). В качестве варьируемых признаков будем использовать 10 морфологических признаков, пол и 3 показателя, определяющих цвет ядовитого секрета и активность отдельных его компонентов. Детальный список признаков представлен в приложении 1 (пример [П5]).

Дискриминантный анализ LDA в этих условиях рассчитывает уравнения двух дискриминирующих плоскостей $Z_1(\boldsymbol{x}) = \boldsymbol{\beta}_1 \boldsymbol{x}$ и $Z_2(\boldsymbol{x}) = \boldsymbol{\beta}_2 \boldsymbol{x}$. Последовательное применение этих решающих правил дает возможность отнести произвольный экзаменуемый объект к одной из трех групп. С использованием функции lda(…) пакета MASS были построены модель LDA1, включающая полный набор переменных, и LDA2 на основе четырех признаков, которые в ходе селекции были признаны информационно значимыми процедурой stepclass(…):

$Z_1(\boldsymbol{x}) = 0.091\ S.cd. - 0.050\ ПА - 0.021\ L.AMO - 5.351\ Col$;

$Z_2(\boldsymbol{x}) = 0.193\ S.cd. - 0.104\ ПА + 0.029\ L.AMO + 1.133\ Col$,

где *S.cd.* - количество пар подхвостовых щитков, *ПА* - протеолитическая активность яда, *L.AMO* – активность оксидазы *L*-аминокислот, Col – цветность яда от 0 (бесцветный) до 1 (ярко-желтый). Однако, как представлено в табл. 6.9, переход от полной модели к сокращенной в этом примере вызывает существенное увеличение ошибок распознавания.

Если сократить количество используемых переменных до двух, то появляется возможность отобразить границы, разделяющие группы животных, на двухмерной диаграмме. На рис. 6.20 видно, как отличается форма разделяющих поверхностей линейного и квадратичного вариантов дискриминантного анализа.

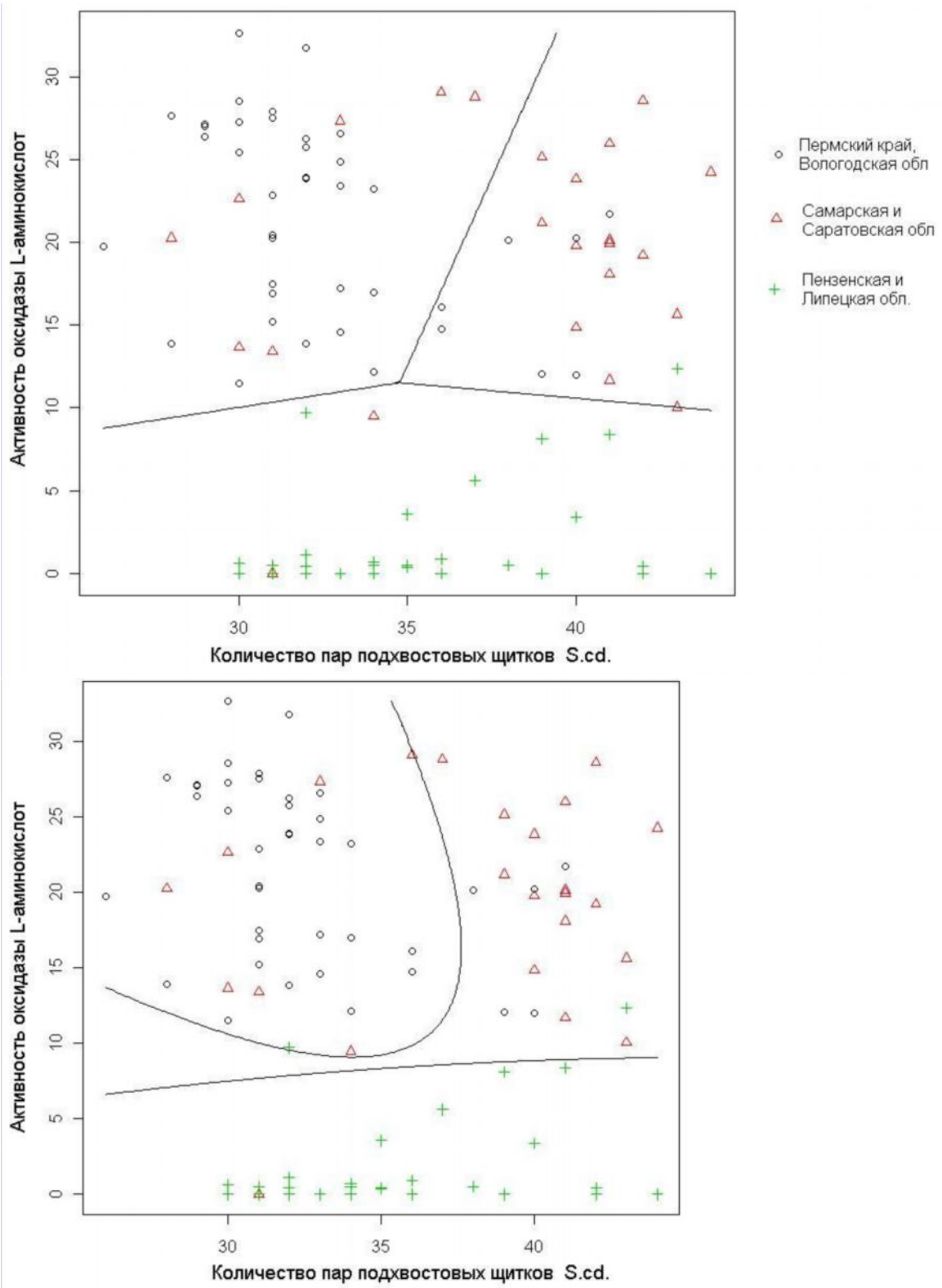


Рис. 6.20. Границы дискриминации трех региональных подпопуляций гадюки обыкновенной в пространстве двух переменных методами LDA (а) и QDA (б)

При использовании метода опорных векторов SVM можно применить стратегию "один против всех", т.е. строятся разделяющие поверхности между каждой группой объектов и всеми остальными и вычисляются оценки принадлежности Z_k по каждому классификатору. Экзаменуемый объект относится к тому классу, которому соответствует максимальное значение Z_k. Если использовать схему "каждый против каждого", то строится $k(k-1)/2$ таких поверхностей. Возможны также стратегии "турнир с

выбыванием", "дихотомия" и др. При этом можно отметить (табл. 6.9) значительное превосходство SVM в качестве распознавания по сравнению с традиционными линейными моделями.

В отличие от уже описанных алгоритмов, в представленных ниже методах многоклассовая классификация реализуется естественным образом, а не сводится к последовательности двухклассовых разбиений.

Метод *k ближайших соседей* (*k* nearest neighbors) или kNN-классификация является простейшим алгоритмом, определяющим разделяющие границы для каждого локального объекта. В варианте 1NN анализируемый объект относится к определенному классу в зависимости от информации о его ближайшем соседе. В варианте kNN каждый объект относится к преобладающему классу ближайших соседей, где *k* - параметр алгоритма. В основе метода kNN лежит *гипотеза компактности*, которая предполагает, что тестируемый объект *d* будет иметь такую же метку класса, как и обучающие объекты в локальной области его окружения.

Решающие правила в методе kNN определяются границами смежных сегментов диаграммы Вороного (Voronoi resselation), разделяющей плоскость на |*n*| выпуклых многоугольников, каждый из которых содержит один и только один объект обучающей выборки – см. рис. 6.21. В *p*-мерных пространствах границы решений состоят уже из сегментов (*p* - 1)-мерных полуплоскостей выпуклых многогранников Вороного.

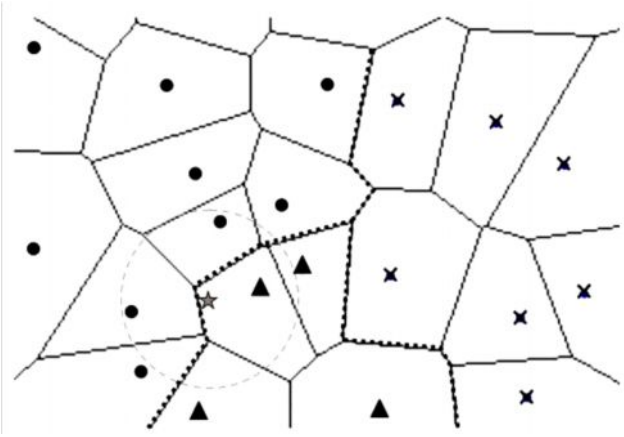


Рис. 6.21. Иллюстрация работы алгоритма *k* ближайших соседей

Алгоритм предсказания строится на основе схемы голосования, т.е. в качестве результата объявляется метка класса-победителя. На рис. 6.20 тестируемый объект "звёздочка" попадает в ячейку объекта класса "треугольников" и при $k = 1$ будет отнесён к этому классу. Однако при $k = 3$ по голосам двух ближайших соседей из трех экзаменуемая точка будет отнесена к классу "кружочков". Вероятностный вариант метода kNN использует для ранжирования предполагаемых классов сумму "голосов" соседей с учетом их весов, в частности, косинусной меры расстояния между тестируемым объектом и каждым из соседей.

Следует отметить, что вариант 1NN обеспечивает 100% правильное распознавание примеров обучающей выборки, однако часто ошибается на неизвестных ему векторах признаков (табл. 6.9). При увеличении $k > 1$ до некоторых пределов качество распознавания на внешнем дополнении начинает возрастать. Наилучшее в смысле классификационной точности значение *k* может быть найдено с использованием кросс-проверки. Для этого по фиксированному значению *k* строится модель *k*-ближайших соседей и оценивается *CV*-ошибка классификации при скользящем контроле. Эти действия повторяются для различных *k* и значение, соответствующее наименьшей ошибке распознавания, принимается как оптимальное. В рассматриваемом примере наилучший прогноз основан на количестве соседей $k = 5$ и $k = 8$ – рис. 6.22.

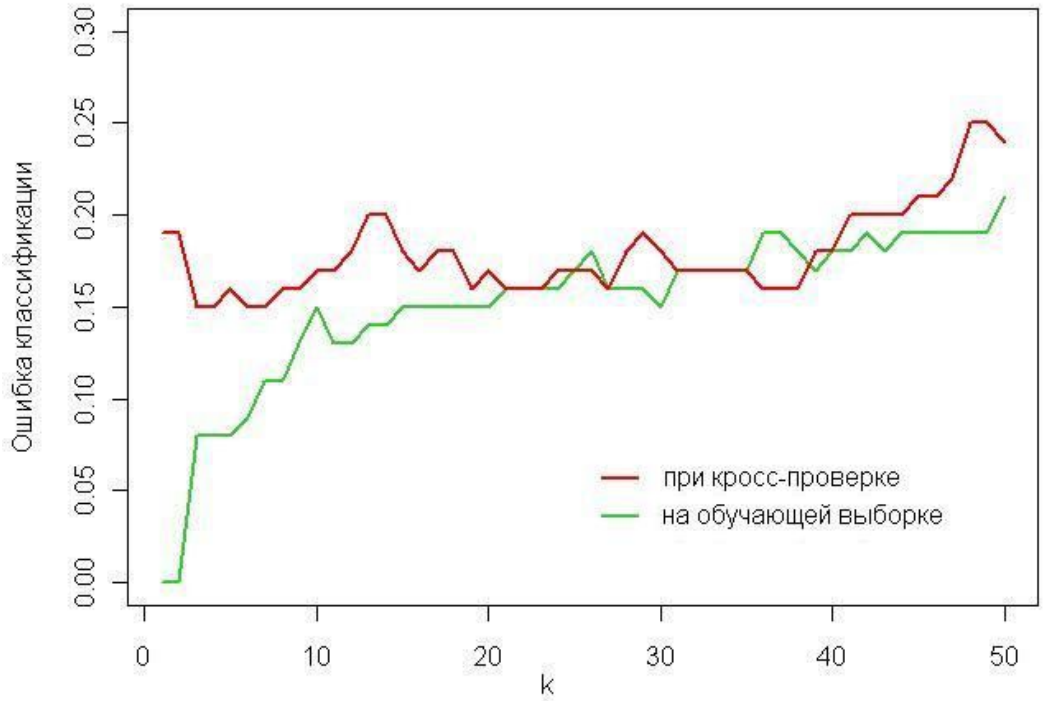
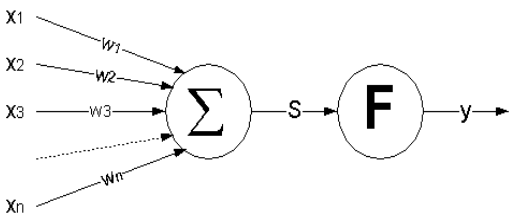


Рис. 6.22. Нахождение оптимального числа *k* ближайших соседей при классификации популяций гадюки

Нейроинформатика, бурное развитие которой пришлось на 80-е годы прошлого столетия, в настоящее время стала определяющей парадигмой создания сложных систем распознавания во многих областях науки и техники. Искусственные нейронные сети (ИНС – Уоссермен, 1992; Шитиков и др., 2005) конструируются из достаточно простых элементов – *формальных нейронов*, напоминающих свой биологический прототип.



На n входов каждого нейрона (синапсы) поступают значения переменных $x_1 \ldots x_n$, а на выходе генерируется результирующая величина $y = F(S)$, где S – взвешенная сумма входных сигналов, F – *функция активации* нейрона, преобразующая S в выходной сигнал.

Для суммации входных сигналов используется выражение $S = \sum_{i=1}^{n} w_i x_i - T$, где T – порог нейрона, $w_1..w_n$ – *веса* синапсов, которые могут быть как тормозящими, так и усиливающими. Вид функции активации F может иметь различное математическое выражение, выбор которого определяется характером решаемых задач. Если, например, использовать сигмоид $F(S) = 1/(1+e^{-cS})$, то нейрон реализует логистическую регрессию.

В основе нейросетевого подхода лежит идея построения распознающего устройства из большого числа параллельно работающих простых элементов (т.е. описанных выше формальных нейронов), которые функционируют независимо друг от друга и связаны между собой однонаправленными каналами передачи информации. Хотя каждый отдельный нейрон моделируется довольно простой функцией регрессии, совокупная сложность модели, гибкость ее функционирования и другие важнейшие качества определяются структурой связей и многоуровневой иерархией всей сети.

В настоящее время описано много разновидностей нейронных сетей (Горбань и др., 1998), характерных для различных типов задач. Рассмотрим возможности обучения ИНС на примере трехслойного персептрона с прямым распространением информации, представленного на рис. 6.23 и включающего 6 нейронов в промежуточном (скрытом)

слое. На входы сети подаются нормированные значения 13 морфологических показателей и параметров активности ядовитого секрета, а на выходе моделируется значение отклика Y в виде метки географического региона, где предположительно обитают различные подпопуляции гадюк.

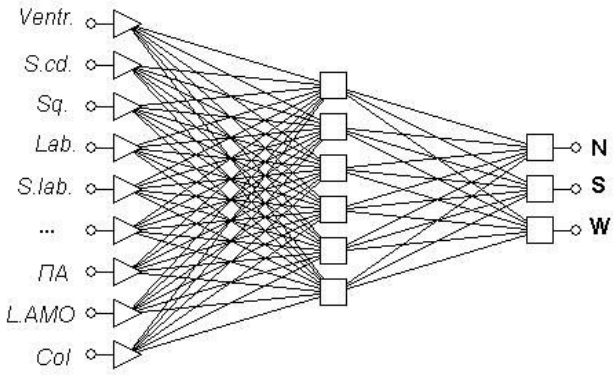


Рис. 6.23. Вид нейронной сети для классификации популяций гадюки

Обучение нейронной сети связано с некоторыми трудностями вычислительного характера: необходимо подобрать число нейронов промежуточного слоя, значение параметра ослабления λ и другие параметры (Venables, Ripley, 2002), причем сам процесс настройки имеет нестационарный характер и требует проведения некоторого числа повторностей. Эффективность распознавания обученной сети (см. табл. 6.9), как правило, высока, однако содержательная интерпретация модели, полученной на ее основе, практически невозможна.

Таблица 6.9. Доля ошибочной классификации (%) моделей распознавания региона обитания гадюк на обучающей выборке (*Err-S*) и при скользящем контроле (*Err-CV*)

Использованный алгоритм	*Err-S*	*Err-CV*
Линейный дискриминантный анализ LDA (14 признаков)	6	10
Линейный дискриминантный анализ LDA (4 признака)	11	15
Метод опорных векторов SVM	4	4
Метод *k* ближайших соседей kNN (*k* = 1)	0	19
Метод *k* ближайших соседей kNN (*k* = 5)	8	15
Персептрон с одним скрытым слоем из 6 нейронов	0.3	–

К разделу 6.8:

```
VIP <-read.delim("Змеи.txt"); attach(VIP)
library(klaR)    #  -------- Дискриминантный  анализ
lda.all <- lda(Region ~ ., VIP); Out_CTab(lda.all, VIP$Region)  # Функцию См. Раздел 7.8
stepclass(Region ~ ., data=VIP, method="lda")
lda.step <- lda(Region ~ S.cd.+ ПА + L.AMO + Col,  VIP) ; Out_CTab(lda.step, VIP$Region)
#  Функция визуализации биплота с разделяющими границами
predplot <- function(object, x, main = "", len = 200, ...)  {
    xp <- seq(min(x[,1]), max(x[,1]), length=len)
    yp <- seq(min(x[,2]), max(x[,2]), length=len)
    grid <- expand.grid(xp, yp)  ;  colnames(grid) <- colnames(x)[-3]
    Z <- predict(object, grid, ...) ;     zp <- as.numeric(Z$class)
    zp <- Z$post[,3] - pmax(Z$post[,2], Z$post[,1])
    plot(x[,1], x[,2], col = x[,3], pch = x[,3], main = main)
    contour(xp, yp, matrix(zp, len), add = T, levels = 0, drawlabels = FALSE)
    zp <- Z$post[,1] - pmax(Z$post[,2], Z$post[,3])
    contour(xp, yp, matrix(zp, len), add = T, levels = 0, drawlabels = FALSE)  }
#  Вывод диаграмм на рис. 6.20
lda.2 <- lda(Region ~ S.cd.+ L.AMO,  VIP) ; qda.2 <- qda(Region ~ S.cd.+ L.AMO,  VIP)
X2 <- data.frame(VIP[,c(4,14)], as.numeric(VIP$Region))
predplot (lda.2, X2)  ;predplot (qda.2, X2)
```

```
library(e1071)    #   -------- Метод опорных векторов SVM
svm.cv <- svm(formula = Region ~ .,  data = VIP, cross=10, kernel = "linear")
table(Факт=VIP$Region, Прогноз=predict(svm.cv))  ; mean(predict(svm.cv) != VIP$Region )
n <- nrow(VIP) ; Err_S  <- 0 ; for(i in 1:n)   #   Выполнение кросс-проверки
   {svm.temp <- svm(formula = Region ~ .,  data = VIP[-1,], cross=10, kernel = "linear")
   if (predict(svm.temp, newdata=VIP[i,]) != VIP$Region[i]) Err_S <- Err_S +1 } ; Err_S/n }
library(kknn)   #   -------- Метод к-ближайших соседей  kNN
gen.err.kknn <- numeric(50) ; mycv.err.kknn <- numeric(50) ; n <- nrow(VIP)
for (k.val in 1:50)  {  #  Рассматриваем число возможных соседей от 1 до 50
  pred.kknn <- kknn(Region ~ ., VIP,train=VIP,test=VIP,k=k.val,kernel="rectangular")
  gen.err.kknn[k.val] <- mean(pred.kknn$fit != VIP$Region)
  for (i in 1:n)   {
     pred.kknn <- kknn(Region ~ ., train=VIP[-i,],test=VIP[i,], k=k.val,kernel="rectangular")
     mycv.err.kknn[k.val] <- mycv.err.kknn[k.val] + (pred.kknn$fit != VIP$Region[i])   }
}  ;  mycv.err.kknn <- mycv.err.kknn/n
plot(c(1:50),gen.err.kknn,type="l",xlab='k',
      ylab='Ошибка классификации',ylim=c(0,0.30),col="limegreen", lwd=2)
points(c(1:50),mycv.err.kknn,type="l",col="red", lwd=2)
#  Аналогичный результат получаем при использовании функции  train.kknn(…)
train.kknn(Region ~ ., VIP, kmax = 50, kernel="rectangular")
library(nnet)   #  -------  Нейронные сети
R <- VIP$Region ; VIP.scale <- scale(VIP[,3:15]) ; nreps<-10 ; gen.err.nnet <- numeric(10)
# подбираем число нейронов s скрытого слоя, каждый раз выполняя по 10 повторностей
for (s in 1:10)  {  for (k in 1:nreps) {
   vip.nnet <- nnet(R ~ ., data = VIP.scale, size=s, trace=F) ; vip.p<-predict(vip.nnet)
   gen.err.nnet[s] <- gen.err.nnet[s]+100*sum(as.numeric(R)!=max.col(vip.p))/length(R)
  }  } ; (gen.err.nnet <- gen.err.nnet/nreps)
```

6.9. Самоорганизующиеся карты Кохонена

Отображение структуры данных может быть реализовано с использованием нейронных сетей особого типа – так называемых самоорганизующиеся структур. Формируемые при этом самоорганизующиеся карты (SOM – Self Organizing Maps) стали мощным аналитическим инструментом, объединяющим в себе две основные парадигмы анализа – кластеризацию и проецирование, т.е. визуализацию многомерных данных на плоскости.

В сетях SOM (см. рис. 6.24) на входы двухмерной решетки нейронов **N** подается многомерный образ данных **S**, состоящий из векторов выборки наблюдений. С каждым узлом (нейроном) решетки, которая может иметь прямоугольную или гексагональную форму, ассоциируются базисные векторы $m_i = (\mu_{1i}, \mu_{2i}, \ldots, \mu_{1n})$, определяющие потенциальные центры кластеров. Т. Кохонен (Kohonen, 1982) предложил модификацию алгоритма Хебба соревновательного обучения "без учителя", в результате чего пропорциональный вклад стали получать не только нейроны-победители, но и ближайшие их соседи, расположенные в окрестности $\boldsymbol{R}$. Вследствие этого выходные сигналы нейронов стали коррелировать с положением прототипов в многомерном пространстве входов сети, т.е. близким нейронам стали соответствовать близкие значения входов $\boldsymbol{S}$.

"Проекционный экран", на который выводятся результаты обучения SOM, имеет вид упорядоченной структуры, свойства которой плавно меняются на двумерной сетке координат, образуя как бы краткое резюме исходных данных. Степень регулярности топографии сети и метрически близких векторов исходных данных определено здесь на нескольких уровнях.

° сохранение *топологии* (на рис. 6. экстремальные точки 1, 5, 21, 25 поместились в углах экрана, а сгущение в центре многомерного облака распределилось по центральным ячейкам);

° сохранение *порядка* расстояний между парами точек данных и соответствующими парами нейронов, на которые эти точки отображаются;

° сохранение *метрических свойств*, т.е. еще более строгое понимание подобия, основанное на прямом вычислении численных метрических отношений между прототипами и их отображениями.

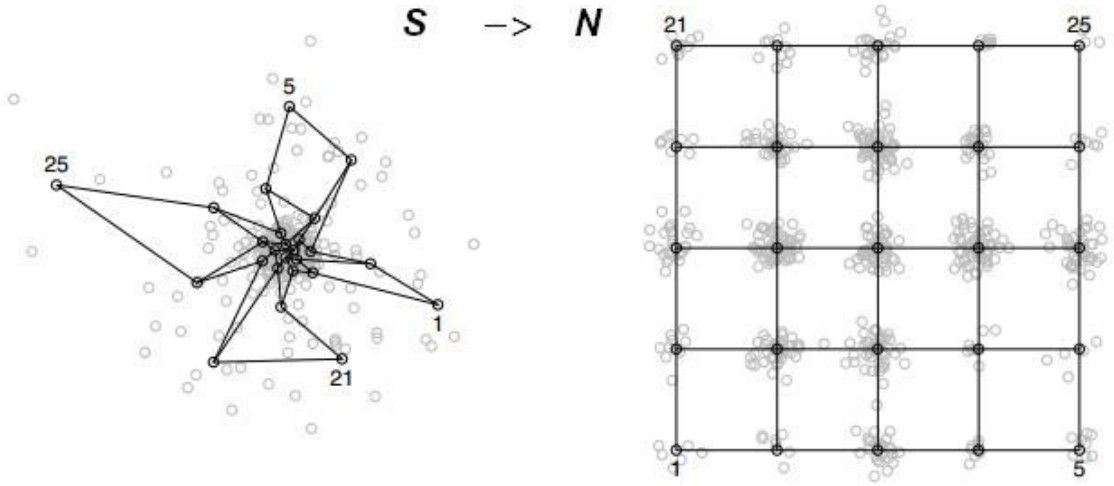


Рис. 6.24. Иллюстрация процесса отображения SOM: многомерное облако точек наблюдений ***S*** проецируется на экран из 25 нейронов сети ***N*** (Wehrens, 2011)

Можно выделить несколько существенных отличий SOM от других ординационных методов. Например, PCA формирует оси главных компонент, основываясь на масштабировании дисперсий, в результате чего большая группа точек оказывается близко к центру и трудно опознается (см. рис. 6.4). Использование различных методов построения дендрограмм в силу своей одномерности также не позволяет отобразить всю структуру “взаимоотношений” классов. В этих условиях нейронные сети Кохонена в силу своей адаптивности и самоорганизации не требуют предварительной калибровки данных и позволяют выявить внутреннюю структуру объектов с учетом всей совокупности выборочных точек.

Традиционно SOM рассматривается как эмпирический алгоритм, а выводы (в первую очередь, качественные) о структуре данных делаются на основе визуального анализа полученной карты. Математический аппарат SOM также является трудным для описания, поскольку алгоритм определяется не в терминах функции ошибок, а через многочисленные эвристические правила итеративного обновления весов нейронов. Достаточно подробное русскоязычное описание техники расчетов и формул, применяемых для сглаживания и упорядочения базисных векторов при обучении, представлено, например, в руководствах к программе ScanEx IMAGE Processor v3.6 (http://www.scanex.ru/ru/software).

Рассмотрим в качестве примера задачу районирования Куйбышевского водохранилища [П1] по результатам многолетнего мониторинга. Сформируем выборку из 6 показателей: средних (за июнь-август) значений биомассы каляноид (CAL) и диатомовых водорослей (DIA), а также численности бактерий (BAK), при сопутствующих внешних факторах – прозрачности воды (PRO), содержании общего азота (Nsu) и температуры (TEM) в придонном слое. Всего исходная таблица содержит 305 строк данных, полученных в разные годы на 15 станциях наблюдения по всей акватории водохранилища.

Первой проблемой, с которой мы сталкиваемся – наличие пропусков в данных. В статистической среде R существует большое число функций, заполняющих пропуски в многомерных таблицах с использованием продвинутых методов, реализующих различные статистические модели аппроксимации: mitools(…), mice(…), mix(…) из одноименных пакетов, aregImpute(…) и transcan(…) из пакета Hmisc и др. Однако в учебных целях

используем свою простейшую функцию, заполняющую пропуски случайными значениями этого же показателя из подмножества зарегистрированных наблюдений для каждой станции. Для удобства последующего анализа в табл. 6.10 приведем средние выбоочные показатели по плесам.

Таблица 6.10. Средние выборочные показатели по плесам Куйбышевского водохранилища

Участки водохранилища	Каляноиды мг/м^3	Диатомовые, г/м^3	Бактерии, млн.кл./мл	Прозрачность, см	Азот общий, мг/л	Температура, град.
Волжский	51.9	9.52	1.58	84.1	0.423	19.4
Волго-Камский	65.1	10.2	1.71	76.3	0.396	18.9
Тетюшинский с Ундорским	83.4	1.50	1.47	101.9	0.422	18.3
Ульяновский	142.6	2.17	1.38	106.2	0.497	17.8
Приплотинный	83.4	1.71	1.10	139.2	0.455	17.3

Обучение сети для этого примера выполним с использованием функции som(…) пакета kohonen, задав гексагональную решетку проекционного экрана 5×6, т.е. 305 точек наблюдений будут "самоорганизовываться" на 30 узлах – потенциальных центрах кластеризации. По завершении итерационного процесса нам становится доступным для визуализации следующий комплект карт:

- "codes" – показывается распределение по решетке соотношение долей участия отдельных исходных переменных (см. рис. 6.25а);
- "counts" – представлено число исходных объектов в каждом узле сети;
- "mapping" – показываются координаты исходных объектов на сформированной карте (см. рис. 6.25б);
- "property", "quality", "dist.neighbours" – визуализируется цветом различной окраски целый набор свойств каждого узла: меры парных или средних расстояний между нейронами, доли участия отдельных исходных переменных и т.д.

При анализе полученных карт можно усмотреть, что кластеры в нижнем левом углу решетки связаны с высокой численностью бактерий и содержанием диатомовых водорослей, а также с повышенной температурой воды в придонном слое на мелководьях. Их "заселяют" в основном наблюдения, сделанные на Волжском и Волго-Камском участках. Для верхнего левого угла экрана характерно присутствие станций Приплотинного плеса с высокой прозрачностью и большим содержанием азота. Некоторые точки наблюдений Ундорского участка с повышенной биомассой каляноид нашли свое место в верхнем правом углу. Наконец, многообразие природных характеристик Ульяновского плеса, где глубоководные русловые участки сочетаются с мелководными заливами, вызвало неакцентированную "размытость" размещения синих точек почти по всей решетке.

К разделу 6.9:

```
#  Самоорганизующиеся карты Кохонена
# Функция случайного заполнения пропущенных данных в произвольном векторе
random.imp <- function (a){
    missing <- is.na(a) ; n.missing <- sum(missing) ; a.obs <- a[!missing] ; imputed <- a
    imputed[missing] <- sample (a.obs, n.missing, replace=TRUE) ; return (imputed) }
# Инициализация данных и заполнение пропусков
library(xlsReadWrite) ; PT <- read.xls("Куйб.xls", sheet = 2, rowNames=FALSE)
Stan.f <- as.factor(PT[,2]) ; X <- as.data.frame(PT[,4:9])
Stan.list <-unique(Stan.f) ; XP <- X
for (i in 1:length(Stan.list)) {
    Ind <- which(Stan.f==Stan.list[i]); XP[Ind,] <- apply(X[Ind,], 2, random.imp) }
by(XP, Pl.f, function(x) apply(x, 2, mean))    #  Вывод таблицы средних по плесам
```

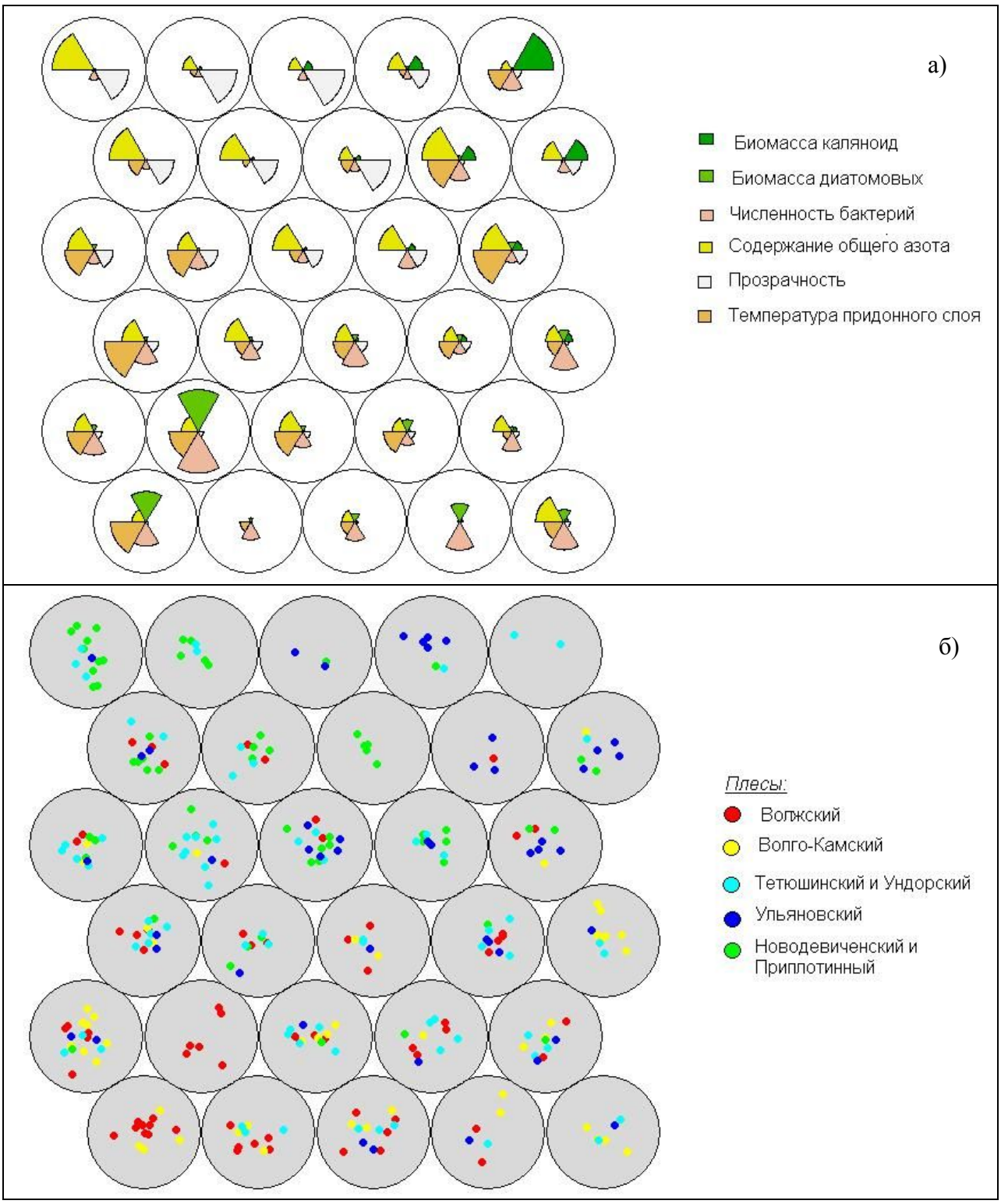


Рис. 6.25. Самоорганизующиеся карты типа "codes" (а) и "mapping" (б), проецирующие 305 наблюдений 6 показателей на акватории Куйбышевского водохранилища

```
library(kohonen)                          #  Обучение SOM и вывод карт на экран
PT.som <- som(data = as.matrix(XP), grid = somgrid(5, 6, "hexagonal"))
plot(PT.som, type = "codes") ; plot(PT.som, type = "counts")
plot(PT.som, type = "dist.neighbours") ; Pl.f <- as.factor(PT[,1])
plot(PT.som, type = "mapping", col =as.integer(Pl.f), pch = 16, bgcol = gray(0.85))
```

7. АНАЛИЗ ПРОСТРАНСТВЕННО-ВРЕМЕННОЙ ДИНАМИКИ И БАЙЕСОВСКИЕ МЕТОДЫ

7.1. Декомпозиция временных рядов и выделение тренда

В предыдущих разделах при разработке моделей регрессии или выполнении многомерного анализа немаловажным считалось предположение о независимости элементов выборочных рядов. Основными принципами организации наблюдений обычно являются *рандомизация* и *повторность*, что обеспечивает корректную оценку эффектов воздействия факторов и выявление связей между переменными. В случае временных рядов эти принципы нереализуемы и мы имеем простейшую форму зависимых данных, строго упорядоченных на некотором фиксированном интервале

$$x_t, x_{t+\Delta t}, x_{t+2\Delta t}\ldots,$$

где t – начальный момент времени, Δt – периодичность отбора проб. Если Δt = const, то обычно отсчеты времени представляются натуральным рядом чисел $t = 1, 2, \ldots, n$.

При анализе временных рядов предполагается, что последовательность x_t является порождением случайного процесса, т.е. конкретное наблюдение рассматривается как одно из множества возможных значений, определяемых распределением вероятностей $p(x_1, ..., x_n)$. Этот процесс является стационарным, если он находится в определенном смысле в статистическом равновесии и его свойства не зависят от времени. Под влиянием факторов различной природы динамический ряд может оказаться нестационарным и одной из основных задач анализа является выявление источников такой нестационарности (Cowpertwait, Metcalfe, 2009).

Для интерпретации последующих расчетов в качестве экологических иллюстраций будем использовать временные ряды (пример [П1]), имеющие следующие условные обозначения и характеристики:

- СКОРОСТЬ и ПОВТОР – среднемесячная скорость (м/сек) и повторяемость северного ветра (%) по данным одного из метеопостов в районе г. Тольятти за период с 1961 по 1988 г.; всего 336 измерений;
- РАСХОД – суммарный расход воды в Куйбышевском водохранилище по плотине Волжской ГЭС, км3/мес. по данным за период с 1957 по 1988 г.; всего 384 измерений.

При анализе временных рядов выделяются две ключевых концепции: выделение постоянного (условно-монотонного или унимодального) тренда и оценка корреляции между равноудаленными членами ряда, связанной с синхронностью периодических колебаний. Для этого выполняется разложение временного ряда на составляющие (компоненты), которые с экологической точки зрения несут разную содержательную нагрузку. Аддитивная модель декомпозиции имеет вид:

$$x_t = m_t + s_t + z_t, \quad \text{где}$$

- m_t – тренд (trend) или детерминированная компонента, выражающая основную направленную тенденцию изменчивости динамического ряда;
- s_t – периодическая составляющая, моделирующая сезонные, календарные или иные циклические вариации исследуемого объекта;
- z_t – стохастическая компонента (или остатки), представляющая собой последовательность скоррелированных случайных значений с нулевым средним.

Выделение отдельных компонентов декомпозиции ряда СКОРОСТЬ (рис. 7.1) осуществлялось следующим образом:

- периодическая составляющая s_t рассчитывалась путем сглаживания сезонных серий локальными полиномами; доля вариации s_t, оцениваемая как соотношение межквартильных размахов циклического и исходного рядов, составила d_{IQR} = 27.6%;
- функция тренда m_t оценивалась с использованием модели локальной регрессии, аппроксимирующей ряд, полученный после исключения из него циклической компоненты s_t, (d_{IQR} = 63.1%); если функцию тренда выделить непосредственно из исходного ряда, то доля объясняемой вариации составляет только d_{IQR} = 51.1%;

° после вычитания из исходного ряда тренда и периодической составляющей был образован ряд остатков $z_t = x_t - m_t - s_t$.

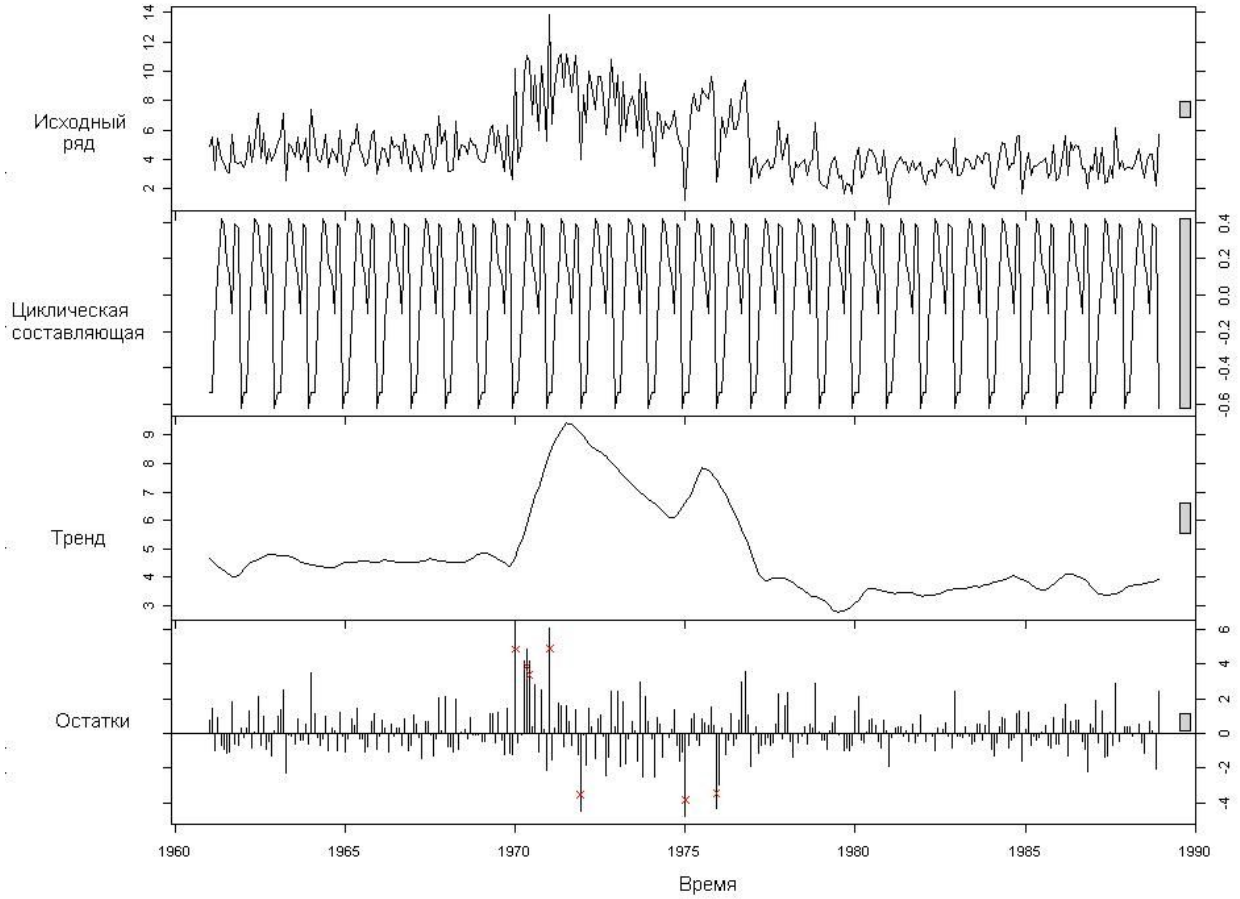


Рис. 7.1. Декомпозиция ряда СКОРОСТЬ с выделением тренда и циклической составляющей; красными крестиками отмечены предполагаемые выбросы

В общем случае существуют различные категории методов разложения временных рядов на компоненты с выделением тренда и периодической составляющей. Первая группа использует модели множественные регрессии с постоянными параметрами и факторами, являющимися функциями времени. Ко второй категории можно отнести непараметрические модели сглаживания с коэффициентами, динамически вычисляемыми в каждой точке диапазона наблюдений независимой переменой. Третья группа методов основана на применении линейных фильтров и базовой моделью является комбинированная модель авторегрессии-интегрированного скользящего среднего (АРИСС-ARIMA). Наконец, можно упомянуть методы, основанные на вейвлет-преобразованиях, алгоритмы "Гусеница" (SSA – Singular Spectrum Analysis), динамический факторный анализ (DFA – Zuur et al., 2003) и многие другие.

Оценивание коэффициентов параметрической регрессии при моделировании тренда обычно происходит по методикам, подробно описанным нами в главах 3-4, поскольку при этом применяются широко распространенные линейная, полиномиальная, экспоненциальная степенная или логистическая функции. Если ряд имеет ярко выраженную периодичность, то для описания изменения среднего уровня переменной используют, как правило, гармоническую функцию $A\cos(2\pi ft + \varphi)$, где A – амплитуда колебаний, f – линейная частота, φ – сдвиг по фазе. Основным преимуществом параметрических моделей является возможность использования большого набора отработанных статистических критериев для идентификации предикторных функций и проверки статистической значимости параметров регрессии.

Различные методы аппроксимации временных рядов с построением регрессионных функций сглаживания (smoothing) также позволяют весьма точно проследить основные тенденции изменчивости изучаемой переменной и наглядно отобразить форму ряда в виде гладкой кривой тренда. В разделе 4.7 нами рассматривались две такие модели: локальная регрессия и кубические сплайны, а здесь мы продолжим развитие этой темы.

Ядерная модель сглаживания отклика x на шкале независимой переменной t основана на предположении, что для оценки значения $\hat{x}(t_0)$ в точке t_0 наиболее ценными являются наблюдения, находящиеся в окрестности $[t_0 - h, t_0 + h]$, т.е. в "окне" шириной $2h$. Тогда для функции регрессии среднего значения переменной $\hat{x}$ в окне $2h$ можно воспользоваться оценкой Надарайа-Уотсона (Nadaraya, Watson):

$$\hat{x}(t_0) = \sum_{i=1}^{n} x_i w(t_i, t_0, h) / \sum_{i=1}^{n} w(t_i, t_0, h), \text{ где } w(t_i, t_0, h) = K(u)/h, \quad u = (t_i - t_0)/h,$$

$K(u)$ – некоторая симметричная ядерная функция, интегрируемая на всем интервале варьирования данных. Если взять интеграл от $K(u)$ по всей области определения, которой может быть отрезок (обычно [-1, 1]) или вся числовая ось, то $\int K(u)du = 1$ – см. рис. 7.2а.

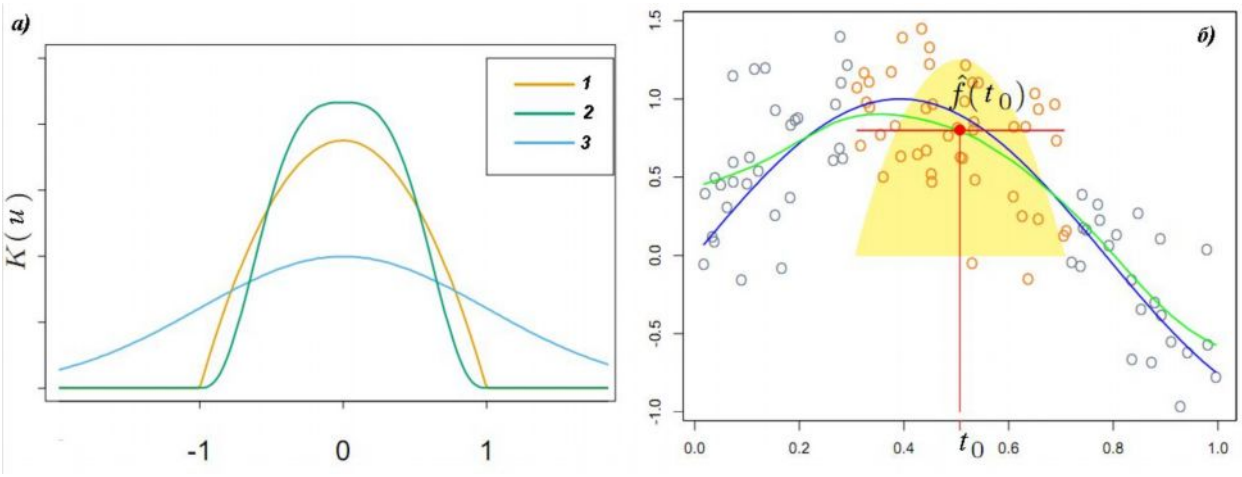


Рис. 7.2. Слева (*а*) - сравнение трех популярных ядерных функций:
1 - ядро Епанечникова $K(u) = 3(1 - u^2)/4$; ***2*** – "трижды кубическая функция" $(1-|u|^3)^3$;
3 – функция нормального гауссового ядра $K(u) = (e^{-u^2/2})/\sqrt{2\pi}$, которая, в отличие от предыдущих функций, не принимает значение 0 за пределами интервала [-1, 1];
Справа (*б*) – пример сглаживания точек в окне, образованным ядром Епанечникова (область, закрашенная желтым цветом) относительно текущего значения t_0; черным показана теоретическая кривая, зеленым – кривая ядерного сглаживания, в подстройке весов которой каждый раз участвуют только точки, отмеченные красным цветом.

Регрессионная кривая оценивается последовательно по мере передвижения окна фиксированной ширины (рис. 7.2б) слева направо по шкале t, в результате чего новая порция наблюдений включается в расчет взамен исключаемых старых. Значения $\hat{x}(t_0)$, полученные сглаживанием в точке t_0, оцениваются с учетом весов w, причем наибольшие веса получают наблюдения, находящиеся ближе к t_0. Воспользовавшись такой моделью, мы проделываем своеобразный "ядерный трюк": нам нет уже необходимости задумываться над выбором базисных функций регрессии, либо остерегаться, что не подтвердятся те или иные исходные предположения анализа (Анатольев, 2009). Нам достаточно выбрать ширину окна h и вид $K(u)$ ядерного регрессора (если нет иных предпочтений, разумно воспользоваться гауссовым ядром).

Роль параметра ширины окна h чрезвычайно велика – см. пример аппроксимации ряда СКОРОСТЬ на рис. 7.3 с использованием гауссового ядра. Если h слишком велик, то в расчетах будет задействована лишняя информация, не характерная для моделируемого поддиапазона данных, что приводит к явлению сверхсглаженности: кривая вырождается в прямую и перестает отслеживать тенденцию изменения наблюдений. Если же h слишком

мал, то модель начинает следовать за случайными флуктуациями наблюдений и недосглаженная кривая будет выглядеть чересчур извилистой.

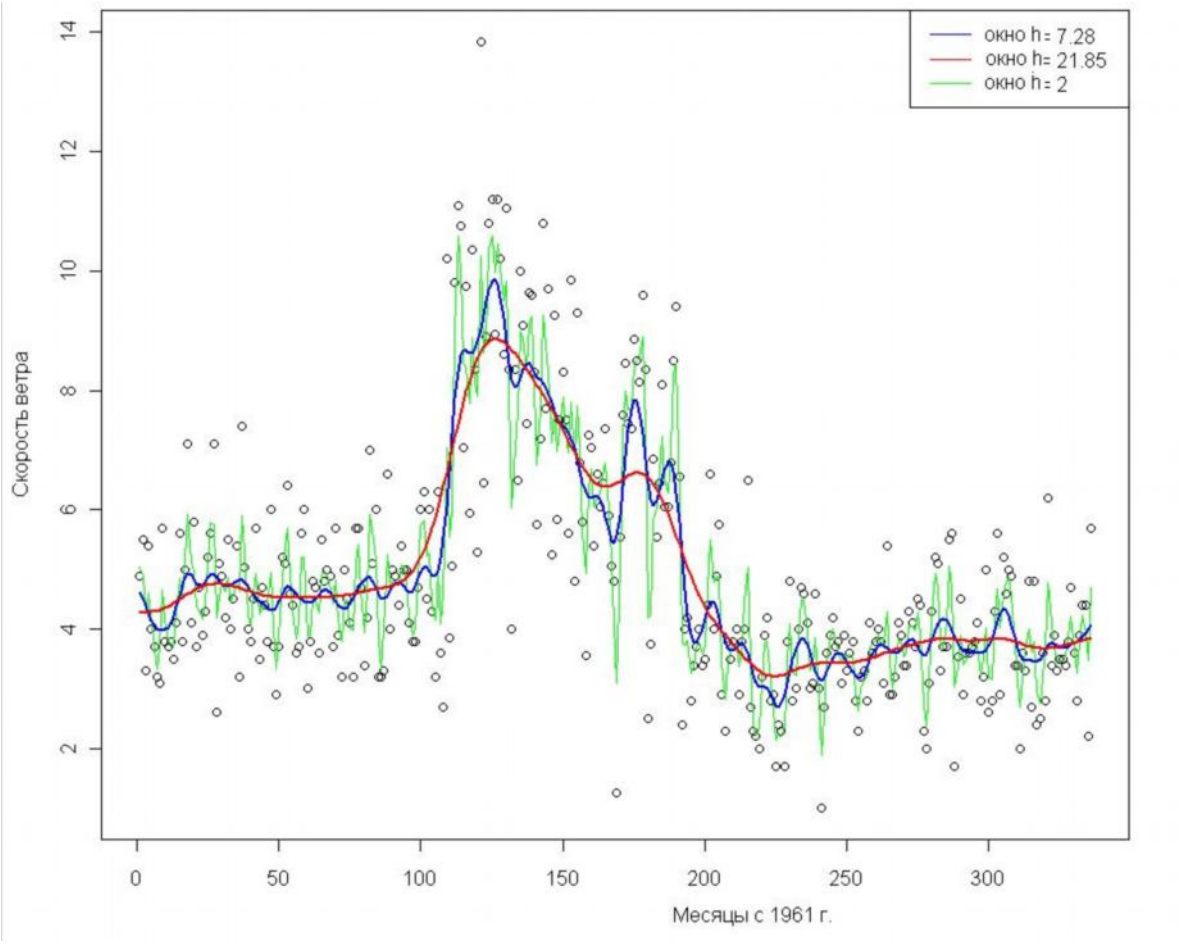


Рис. 7.3. Аппроксимация ряда СКОРОСТЬ ядерной регрессией с различной шириной окна h

Сложность проблемы здесь в том, что если попытаться минимизировать традиционный критерий качества модели – среднеквадратичную ошибку регрессии в точках исходной выборки, то она устремится к нулю и выберет окно минимальной ширины с одним единственным значением (т.е. каждое наблюдение будет объясняться только им самим). Такое сглаживание нас, конечно, не устраивает, поэтому разумной альтернативой поиска оптимального h будут внешние критерии – два варианта функции кросс-проверки (см. выше разделы 3.4, 4.6, 6.1):

$$CV_2(h)=\frac{1}{n}\sum_{i=1}^{n}(y_i-\hat{y}_{-1}(x_i))^2 \quad \text{или} \quad CV_1(h)=\frac{1}{n}\sum_{i=1}^{n}\left|y_i-\hat{y}_{-1}(x_i)\right|$$

где $\hat{y}_{-1}(x_i)$ – оценка Надарайя-Уотсона в точке x_i, основанная на всех наблюдениях, кроме i–го. Оптимальная в смысле процедуры скользящего контроля ширина окна h^{CV} соответствует минимуму ошибки $CV(h)$ на независимом внешнем дополнении данных.

На рис. 7.3 показаны кривые аппроксимации, построенные для набора значений: h = 21.85, доставляющего минимум среднеквадратичной ошибке $CV_2(h)$; h = 7.28, соответствующего минимуму абсолютных средних разностей $CV_1(h)$, и h = 2, как пример недосглаженной кривой.

Другой вариант аппроксимации этого ряда кубическими сплайнами при оптимальном значении параметра сглаживания λ, найденного кросс-проверкой, представлен на рис. 7.4.

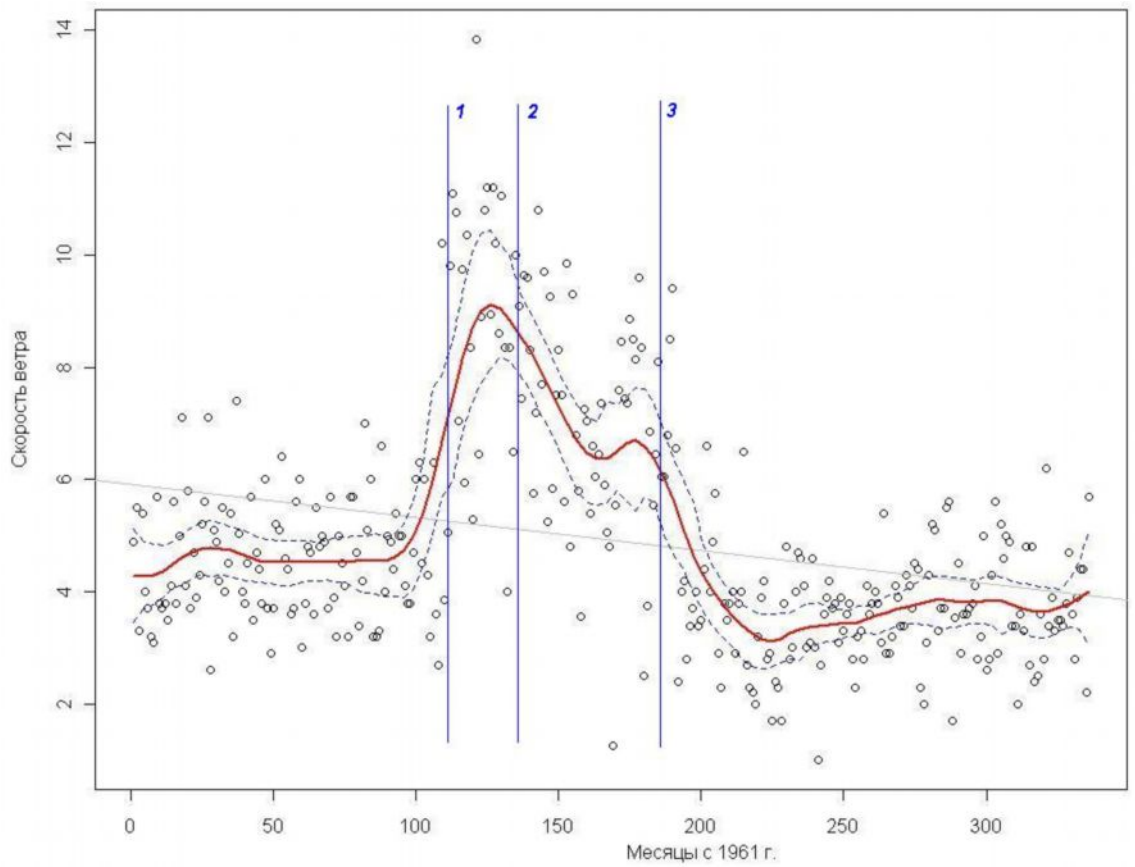


Рис. 7.4. Аппроксимация ряда СКОРОСТЬ сплайном с оптимальным параметром сглаживания; пунктиром показана доверительная область регрессии, вертикалями – критические точки

Выполнив процедуру аппроксимации, недостаточно ограничиться созерцанием полученной картинки со сглаженной кривой. Желательно провести статистический анализ полученной модели с оценкой ее доверительных интервалов, что позволит формализовано подойти к решению следующих важных задач исследования временных рядов:

1. Зависит ли математическое ожидание $m(y|x)$ величины y от временной динамики x в различных частях исследуемой области определения линии регрессии?
2. На каких временных участках можно пренебречь изменчивостью отклика y в зависимости от фактора времени?
3. Какие особые периоды можно трактовать как неустойчивые (бифуркационные), поскольку они сопровождаются наличием экстремумов, перегибов или разрывов первой производной регрессионной кривой?

При использовании непараметрических методов сглаживания часто бывает невозможно оценивать качество подстройки моделей под данные с использованием традиционных статистических критериев (например, оценить значимость коэффициентов по величине t-статистики). Однако доверительные интервалы модели сглаживания несложно найти, если обратиться к процедуре бутстрепа, как это мы делали в разделе 3.4:

- выделим совокупность (например, $L = 100$) опорных временных точек и рассчитаем для них значения скорости ветра по модели сглаживания, построенной по исходным данным (красная кривая на рис. 7.4 соединяет эти точки);
- формируем набор ($B = 1000$) случайных выборок с возвращениями из строк исходной таблицы; по каждой из них строим модель сплайна и рассчитываем 1000 прогнозируемых значений скорости ветра для 100 опорных точек;
- используя распределение значений отклика в вертикальной плоскости каждой из этих точек, находим квантили при $p = 1 - \alpha/2$ и $p = \alpha/2$, $\alpha = 0.5$ и выводим на график линии доверительных интервалов, соединяющие квантильные значения в опорных сечениях (пунктирные линии на рис. 7.4).

Если провести вертикальные секущие плоскости в двух произвольных точках временного ряда и в результате их сравнения доверительные интервалы будут пересекаться, то можно сделать вывод о статистической незначимости различий скорости ветра в этих точках (хотя, как утверждал Р.Фишер, «формально нулевая гипотеза никогда не может быть принята»). Впрочем, если доверительные интервалы не пересекаются, то H_0 также не может быть отвергнута, но по другой причине: мы рассчитали доверительные интервалы огибающих регрессии, а не самой зависимой переменной (см. раздел 3.4).

Большой круг теоретических и практических проблем связан с решением задачи нахождения критических точек, в которых изменяются характеристики ряда (*change-point* analysis). Пусть временная последовательность в одной или нескольких точках перехода τ_1, τ_2, … подвергается резким изменениям в характере распределения своих случайных реализаций. Тогда можно сделать предположение, что весь ряд состоит из сегментов, ограниченных парами τ_i-τ_{i+1}, на каждом из которых имеет место тождественное и независимое распределение F_0, F_1, F_2, … (в общем случае, неизвестное). Проявлением такой нестабильности могут быть отчетливо выраженные скачки (ступени, разрывы) значений среднего, дисперсии, либо какой-либо иной выборочной характеристики.

Г. Росс с соавторами (Ross at al., 2011) предложил следующий алгоритм нахождения количества и локации критических точек τ_i по данным наблюдений. Пусть промежуточный член ряда x_i делит последовательность (x_1, ..., x_n) на две группы. Если у нас нет предположений о законе распределения, то мы можем воспользоваться одним из непараметрических критериев D (Манна-Уитни, Лепажа, Колмогорова-Смирнова, Крамера-Мизеса), чтобы проверить статистическую значимость гипотезы о сдвиге положений или согласии распределений между двумя выборками. Тогда точка перехода τ соответствует индексу i, при котором D принимает максимальное значение при условии отклонения нулевой гипотезы. Сканируя всю последовательность, начиная с x_1 и переходя от значения к значению, можно выделить весь возможный набор τ_1, τ_2, … Для ряда СКОРОСТЬ (рис. 7.4) можно выделить четыре участка, мера положения которых статистически значимо различается по критерию Манна-Уитни, с критическими точками $\tau_1 = 111$, $\tau_2 = 130$, $\tau_3 = 191$ – см. рис. 7.4.

Оценить наличие или отсутствие "горбов" (hump) тенденции временного ряда можно еще одним оригинальным и универсальным способом (Crawley, 2007), который непосредственно не выполняет проверку какой-либо нулевой гипотезы, а основан на моделях иерархической регрессии. При построении деревьев CART (см. раздел 6.4) рекурсивно ищутся такие значения зависимой переменной X, которые делят всю последовательность на подмножества, являющиеся более или менее гомогенными относительно анализируемого признака. Это дает возможность построить характерную кусочно-постоянную линию тренда (см. рис. 7.5а), однородность которой внутри каждого участка связывается с минимум критерия энтропии.

Визуально легко усмотреть, что средняя скорость ветра на втором и третьем участках ($V_{II\text{-}III} = 7.6$ м/с) существенно превышает этот показатель в остальные периоды ($V_{I,IV} = 4.07$ м/с). Но авторитетнее будет выполнить стандартный дисперсионный анализ и показать, что разность между групповыми средними статистически значима при $F_{(3,\ 332)} = 170$ ($p < 0.00001$). Можно также выполнить тест Тьюки и оценить статистическую значимость различий в средней скорости ветра для всех возможных сочетаний пар из четырех выделенных периодов наблюдения – см. рис. 7.5б.

Наконец, еще один подход связан с сегментированием временной последовательности на участки, внутри которых справедлива нулевая гипотеза об отсутствии тренда. Признаком наличия тренда является закономерное изменение во времени статистических моментов первого и второго порядков, что можно проверить с помощью различных критериев, оценивающих случайность членов ряда (т.е. отсутствие взаимосвязи между значениями реализаций наблюдаемой случайной величины и их номерами в выборочной последовательности).

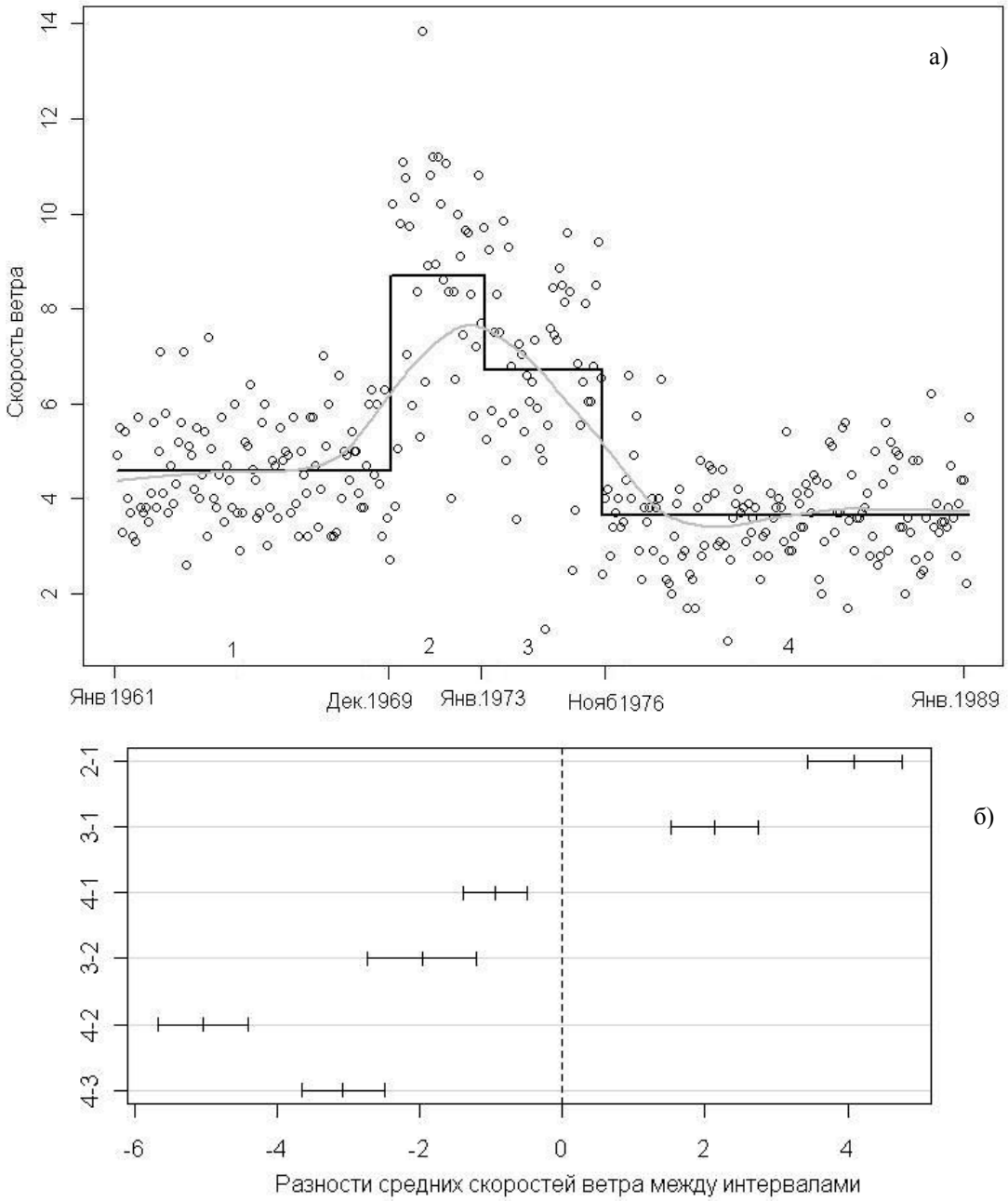


Рис. 7.5. Оценка тренда ряда СКОРОСТЬ с использованием ступенчатой линии; серым цветом показана кривая полиномиального сглаживания (а) и доверительные интервалы разностей скорости ветра между интервалами

Критерий Аббе-Линника Q (Кобзарь, 2006) и его медианная модификация Z (Цейтлин, 2007) проверяют гипотезу H_0: $\mu_t = \mu$, $t = 1, 2, \ldots, n$, что все выборочные значения принадлежат одной генеральной совокупности со средним μ против альтернативы тренда:

$$Q = \frac{\sum_t (x_{t+1} - x_t)^2}{\sum_t (x_t - \bar{x})^2}, \qquad Z = \frac{Me\,|\,x_{t+1} - x_t\,|}{Me\,|\,x_t - \bar{x}\,|},$$

где $\bar{x}$ – выборочное среднее ряда, а Me – медиана выборочных разностей. Уровень достигнутой значимости p, соответствующий вычисленному значению критериальной статистики, может быть определен на основе рандомизационного теста.

При использовании непараметрических ранговых критериев Спирмена и Кокса-Стюарта сравниваются суммы ранговых или знаковых различий временного ряда x_t в хронологической последовательности и ранжированного ряда, отсортированного по

возрастанию x-ов. Оба порядка будут независимы для чисто случайного процесса x_t и скоррелированы при наличии тенденции.

Алгоритм сегментации на бестрендовые участки заключается в сканировании последовательности, начиная с x_1, и для каждого фрагмента временного ряда $\{x_t, x_{t+\Delta}\}$ оценивается статистическая значимость критерия случайности. При отклонении H_0 реализация ряда $x_{t+\Delta-1}$ считается граничным значением предыдущего сегмента, а $x_{t+\Delta}$ – началом нового. В зависимости от используемого критерия для ряда СКОРОСТЬ можно выделить от 8 (критерий Аббе) до 12 (критерий Кокса-Стюарта) участков, в пределах которых нулевая гипотеза о существовании тренда может быть отклонена.

Рассмотрим другой временной ряд РАСХОД, связанный с анализом среднемесячного расхода воды в Куйбышевском водохранилище. Серия наблюдений характеризуется значительной периодичностью: средний расход во время майского паводкового сброса (32.78 км^3) в несколько раз превышает сток в остальные месяцы (от 3 до 5.6 км^3/мес). На фоне этих флуктуаций визуально оценить существование многолетнего тренда чрезвычайно сложно.

Статистическую значимость коэффициента корреляции Спирмена η, оценивающего случайность реализаций всего ряда, можно рассчитать с помощью аппроксимации распределением Стьюдента. Для ряда РАСХОД эта статистика $\eta_{obs} = 0.226$ при $p < 0.0001$, гипотеза H_0: $\eta = 0$ отклоняется и предполагается наличие тренда. Параллельно оценить значимость ранговых различий можно с использованием рандомизационного теста: распределение статистики Спирмена при справедливости нулевой гипотезы со средним $\eta_{ran} = 0.0006$, полученное после 1000 случайных перестановок, не включало эмпирическое значение η_{obs} , т.е. следовательно $p = 0.001$.

Для рассматриваемого ряда РАСХОД трудно получить в явном виде функцию тренда с использованием моделей сглаживания (см. рис. 7.6б). Поэтому, чтобы идентифицировать тип непериодической зависимости на отдельных локальных гомогенных участках, воспользуемся флуктуационным тестом структурной изменчивости параметров линейной регрессии (Kuan, Hornik, 1995).

Идея флуктуационного теста основана на том, что коэффициенты регрессии или их остатки, вычисленные для всего ряда, при отсутствии структурных изменений не должны существенно отличаться от тех же оценок для любых его частных фрагментов. Алгоритм стартует с первого наблюдения и шаг за шагом включает каждую следующую точку, в результате чего рекурсивно пересчитывается суммарная ошибка прогноза на один шаг вперед. Если появляются структурные изменения в тенденции, то ошибка прогноза возрастает и функция CUSUM "накопленных остатков" делает характерный скачок. На графике рис. 7.6а легко увидеть характерный разрыв линейного тренда, приходящийся на март-апрель 1978 г. Напомним, что эта нестационарность расхода связана с созданием новых искусственных водоемов после перекрытия р. Волги у г. Новочебоксарск и р. Камы у г. Набережные Челны.

К разделу 7.1:

```
# Инициализация данных
library(xlsReadWrite) ; WIND <- read.xls("Time_an.xls", sheet = 3, rowNames=FALSE)
OUT <- read.xls("Time_an.xls", sheet = 4, rowNames=FALSE)
#     Создание объектов  "временной ряд"
WIND.speed <- ts(WIND$Скорость, frequency=12, start=c(1961,1))
WIND.repl <- ts(WIND$Повтор, frequency=12, start=c(1961,1))
FLOW <- ts(OUT$Расход, frequency=12, start=c(1957,1)); plot(FLOW)
save(WIND, WIND.speed, WIND.repl, OUT, FLOW, file="time_ser.RData")
#    Вариант декомпозиции ряда СКОРОСТЬ с использованием функции decompose(…)
plot.ts(WIND.speed) ; W.decomp <- decompose(WIND.speed) ; plot(W.decomp)
W.season_adj <- (WIND.speed - W.decomp$seasonal) ; plot(W.season_adj)
#    Вариант декомпозиции с использованием функции stl (…)
stmR <- stl(WIND.speed, s.window = "per", robust = TRUE)
```

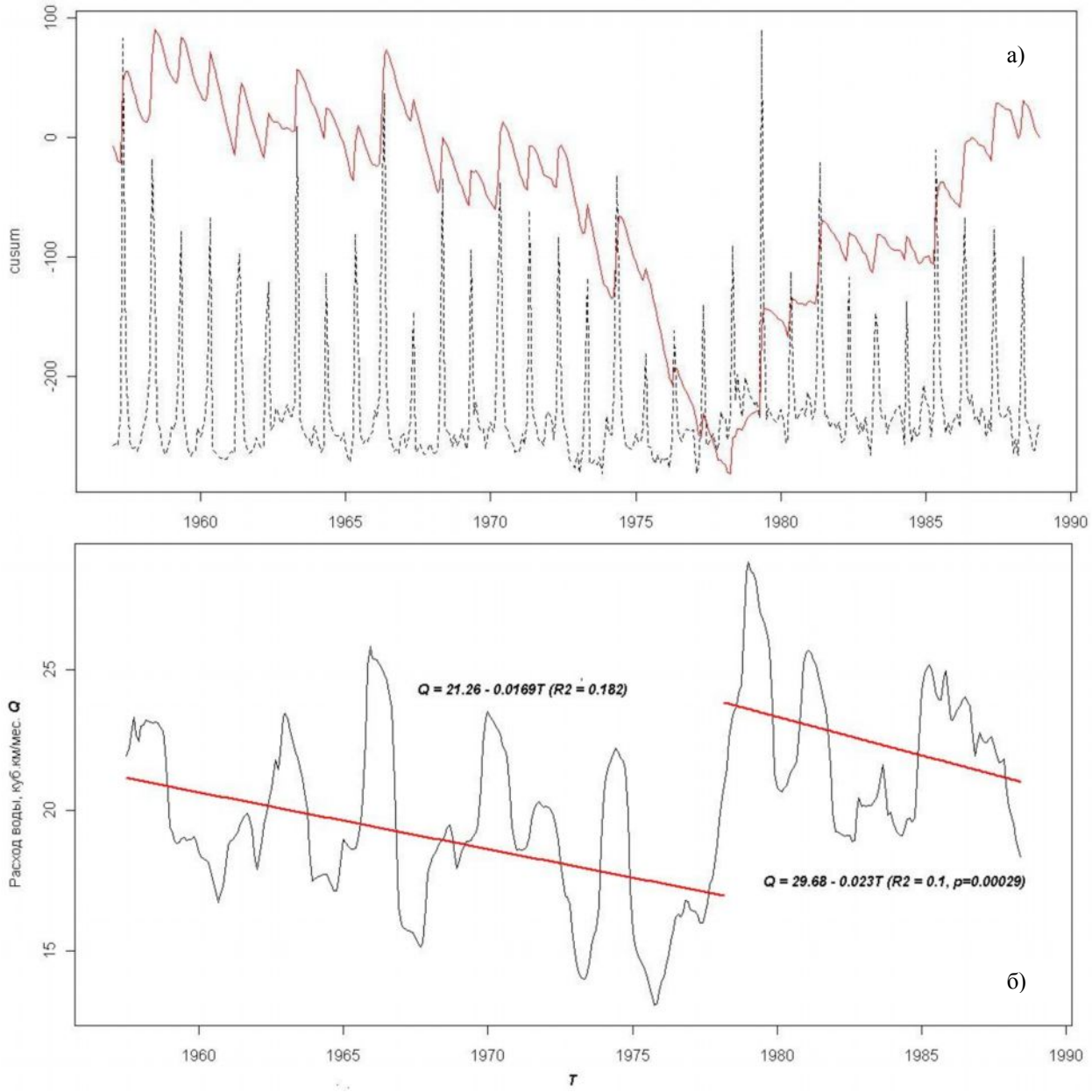


Рис. 7.6. Оценка линейного тренда расхода воды на плотине Куйбышевской ГЭС:
а) график накопленных ошибок прогноза CUSUM при выполнении флуктуационного теста;
б) функция тренда, полученная путем сглаживания ряда локальной моделью LOESS и кусочно-линейной аппроксимацией с разрывом в марте 1978 г.

```
op <- par(mar = c(0,4,0,3), oma = c(5,0,4,0), mfcol = c(4,1)) ; plot(stmR, set.pars=NULL)
 (iO <- which(stmR $ weights < 1e-8)) ; sts <- stmR$time.series
points(time(sts)[iO], 0.8* sts[,"remainder"][iO], pch = 4, col = "red") : par(op)# reset
#    -------------------- Построение моделей сглаживания
sp.Time <- as.numeric(1:nrow(WIND)) ; sp.Speed <- WIND$Скорость
#   Определение функции оценки оптимальной ширины окна кросс-проверкой
bandcrossval<- function(x,y, nstep=20,bmax=0,  L1=TRUE){
    if (bmax==0){bmax<- 1.5*sqrt(var(x))}
    bstep <- bmax/nstep ; n<- length(x); SSE<- c(1:nstep)*0
       for (i in 1:nstep){ for (k in 2:(n-1)){
          xx<- c(x[1:(k-1)], x[(k+1):n]) ; yy<- c(y[1:(k-1)], y[(k+1):n])
          kss<- ksmooth(xx,yy,"normal",ban=(i*bstep),x.points=x[k])
         if(L1==FALSE) { SSE[i]<- SSE[i]+(y[k]-kss$y )^2 }
         if(L1==TRUE) { SSE[i]<- SSE[i]+ abs(y[k]-kss$y )}  }}
    k<- c(1:nstep)*bstep ; return(k[SSE==min(SSE,na.rm = TRUE)])}
#  Сглаживание ядерной функцией и вывод графика на рис. 7.3
bandcrossval(sp.Time, sp.Speed, L1=FALSE) ; bandcrossval(sp.Time, sp.Speed)
```

```
plot(sp.Time,sp.Speed,xlab="Месяцы с 1961 г.",ylab="Скорость ветра")
lines(ksmooth(sp.Time, sp.Speed, "normal", bandwidth=2),col="green")
lines(ksmooth(sp.Time, sp.Speed, "normal", bandwidth=7.28),col="blue",lwd=2)
lines(ksmooth(sp.Time, sp.Speed, "normal", bandwidth=21.85),col="red",lwd=2)
legend ("topright", c("band = 7.28",  "band = 21.85",  "band = 2"),
                                  col = c("blue", "red","green"), lty = 1)
#  Определение функций для расчетов доверительных интервалов сплайна бутстрепом
sp.frame <- data.frame(Time=sp.Time,Seed=WIND$Скорость)
sp.resampler <- function() {     #   Формирование таблицы с перевыборкой
       n <- nrow(sp.frame) ; resample.rows <- sample(1:n,size=n,replace=TRUE)
       return(sp.frame[resample.rows,])   }
sp.spline.estimator <- function(data,m=100) {   #   Расчет прогнозируемых значений сплайна
       fit <- smooth.spline(x=data[,1],y=data[,2],cv=TRUE)
       eval.grid <- seq(from=min(sp.Time),to=max(sp.Time),length.out=m)
       return(predict(fit,x=eval.grid)$y) }
sp.spline.cis <- function(B,alpha,m=100) {  #   Оценка доверительных интервалов
       spline.main <- sp.spline.estimator(sp.frame,m=m)
       spline.boots <- replicate(B,sp.spline.estimator(sp.resampler(),m=m))
       cis.lower <- 2*spline.main - apply(spline.boots,1,quantile,probs=1-alpha/2)
       cis.upper <- 2*spline.main - apply(spline.boots,1,quantile,probs=alpha/2)
       return(list(main.curve=spline.main,lower.ci=cis.lower,upper.ci=cis.upper,
              x=seq(from=min(sp.Time),to=max(sp.Time),length.out=m))) }
#  Сглаживание сплайнами и вывод доверительных интервалов - график на рис. 7.4
sp.cis <- sp.spline.cis(B=1000,alpha=0.05)
plot(sp.Time,sp.Speed,xlab="Месяцы с 1961 г.",ylab="Скорость ветра")
abline(lm(sp.Speed ~ sp.Time),col="grey")
lines(x=sp.cis$x,y=sp.cis$main.curve,col="red",lwd=2)
lines(x=sp.cis$x,y=sp.cis$lower.ci,col="blue",lty=2)
lines(x=sp.cis$x,y=sp.cis$upper.ci,col="blue",lty=2)
# ------------------------------------------------------------------------
# Функция нахождения критических точек изменения последовательности
library (cpm)
cpm.ts <- function(x, cpmType)  {
       # Использует  набор критериев:  cpmType = "Mann-Whitney" , "Mood"
       #                  "Kolmogorov-Smirnov","Cramer-von-Mises", "Lepage" и др.
       cpm <- makeChangePointModel(cpmType=cpmType, ARL0=500) ;  i <- 0
       while (i < length(x)) {
              i <- i + 1 ; cpm <- processObservation(cpm,x[i])
              if (changeDetected(cpm) == TRUE)    {
                     detectiontimes <- c(detectiontimes,i);  Ds <- getStatistics(cpm)
                     tau <- which.max(Ds)
              if (length(changepoints) > 0) {tau <- tau + changepoints[length(changepoints)]}
              changepoints <- c(changepoints,tau) ; cpm <- cpmReset(cpm)
       i <- tau   }  }
  return (list(cpmType=cpmType, changepoints = changepoints, detectiontimes =detectiontimes))
}
cpm.ts(sp.Speed, cpmType="Mann-Whitney") ; cpm.ts(sp.Speed, cpmType="Lepage")
library(tree)    #  Функция построения  ступенчатой линии на рис. 7.5
hump <- function(x,y){  model<-tree(y~x) ; xs<-grep("[0-9]",model[[1]][[5]])
       xv<-as.numeric(substring(model[[1]][[5]][xs],2,10))
       xv<-xv[1:(length(xv)/2)] ; xv<-c(min(x),sort(xv),max(x))
       yv<-model[[1]][[4]][model[[1]][[1]]=="<leaf>"]
       plot(x,y, xlab=deparse(substitute(x)),ylab=deparse(substitute(y)))
       i<-1 ; j<-2 ; k<-1 ; b<-2*length(yv)+1 ; for (a in 1:b){
       lines(c(xv[i],xv[j]),c(yv[k],yv[i]),lwd=2)
       if (a%%2==0){ j<-j+1 ;k<-k+1 }
       else {i<-i+1} }
       return(xv)}
TimeGroup <- hump(sp.Time, sp.Speed) ; lines(lowess(sp.Time, sp.Speed, f=.25),
                                                  col = "grey",lwd=2)
sp.factor <- rep(4, length(sp.Time)) ; sp.factor[sp.Time[sp.Time<TimeGroup[4]]] <- 3
sp.factor[sp.Time[sp.Time<TimeGroup[3]]] <- 2
sp.factor[sp.Time[sp.Time<TimeGroup[2]]] <- 1
mean(sp.Speed[sp.factor==c(1,4)])  ; mean(sp.Speed[sp.factor==c(2,3)])
```

```
sp.factor <- as.factor(sp.factor) ; tapply(sp.Speed,sp.factor,mean)
summary(mod<-aov(sp.Speed~sp.factor)  #   Дисперсионный анализ
#   Выполнение теста Тьюки HSD, вывод графика с доверительными интервалами
(mod.HSD <- TukeyHSD(mod)) ;  plot(mod.HSD)
#  Выполнение сегментации на бестрендовые участки
source("uis.r") ; K.tabl <- uis(sp.Speed)
write.table(K.tabl, file="clipboard", sep="\t", col.names=NA)
# ------------------------------------------------------------------------
#    Декомпозиции ряда РАСХОД с использованием функции decompose(…)
FLOW.d <- decompose(FLOW) ; FLOW.d$figure ; plot(FLOW.d$trend)
#  Тестирование тренда с использованием критерия Спирмена
library(pastecs) ; trend.test(FLOW.d$trend, R=1)   # Оценка p по формулам аппроксимации
#  Тест с рандомизацией
test.Sp <- trend.test(FLOW, R=1000) ; plot(test.Sp) ; test.Sp$p.value
#   Выявление локального тренда
local.trend(FLOW) ; identify(local.trend(FLOW)) ; T=c(1: nrow(FLOW))
fit2 <- lm(FLOW.d$trend[1:255]~T[1:255]) ; summary(fit2)
fit3 <- lm(FLOW.d$trend[255:384]~T[255:384]) ; summary(fit3)
plot(T,FLOW.d$trend, type="l")
points(T[7:255], predict(fit2), type="l", col="red", lwd=2)
points(T[255:378], predict(fit3), type="l", col="red", lwd=2)
```

7.2. Автокорреляция, стационарность и оценка периодичности

Представленные выше методы сглаживания обеспечивают наглядное представление о тренде, но при попытке прогноза на сколько-нибудь ощутимый период часто дают не всегда осмысленные результаты, поскольку не учитывают внутреннюю взаимосвязь членов ряда. Поэтому, прежде чем рассматривать модели прогнозирования, введем понятия автокорреляционной функции и случайного сезонного процесса.

Основным предметом анализа динамики рядов является проверка нулевой гипотезы о том, что изучаемый случайный процесс является *стационарным*, т. е. вероятностные характеристики ряда остаются неизменными. Случайный процесс x_t является стационарным в широком смысле, если его математическое ожидание, дисперсия и автоковариация одинаковы для всех точек последовательности:

$$E(x_t) = \mu = const; \quad E(x_t - \mu)^2 = const; \quad E((x_t - \mu)(x_{t-k} - \mu)) = const.$$

Разработано достаточно большое количество критериев стационарности, различия которых связаны с тем смыслом, который вкладывается в альтернативу проверяемой гипотезы: наличие тренда, его тип, наличие периодичности или что-то другое.

Поскольку члены временного ряда являются взаимозависимыми, для полной характеристики динамического случайного процесса недостаточно математического ожидания и дисперсии. Степень тесноты статистической связи между наблюдениями, "разнесенными" во времени на k единиц, определяется коэффициентом автокорреляции

$$\rho_k = E[(x_t - \mu)(x_{t-k} - \mu)] / \sigma^2 ,$$

где μ и σ^2 – математическое ожидание и дисперсия анализируемой случайной величины.

Выборочная оценка коэффициента автокорреляции имеет вид: $$r_k = \frac{\sum_{t=k+1}^{n}(x_t - \bar{x})(x_{t-k} - \bar{x})}{\sum_{t=1}^{n}(x_t - \bar{x})^2},$$

где n – длина временного ряда. Величина r_k изменяется в зависимости от лага k: например, r_1 оценивает линейную зависимость между парами последовательно выполненных измерений x_t и x_{t+1}, r_2 – между x_t и x_{t+2} и т.д. Автокорреляционной функцией (АКФ) называют последовательность автокорреляций $\{r_k\}$, $k = -\infty, \ldots, +\infty$ при $r_0 = 1$ и $r_k = r_{-k}$.

На рис. 7.7 представлена коррелограмма ряда ПОВТОР среднемесячной повторяемости северного ветра (% дней от общего их количества с оценкой по одному из 8 возможных румбов). Можно отметить умеренную, хотя и значимую ежегодную

цикличность этого метеорологического явления: наиболее тесная прямая связь наблюдается с периодичностью в 1 год, а обратная – через каждые полгода. Аналогичный вывод можно сделать также на основе графика средних месячных значений повторяемости северного ветра за анализируемый период (рис. 7.8).

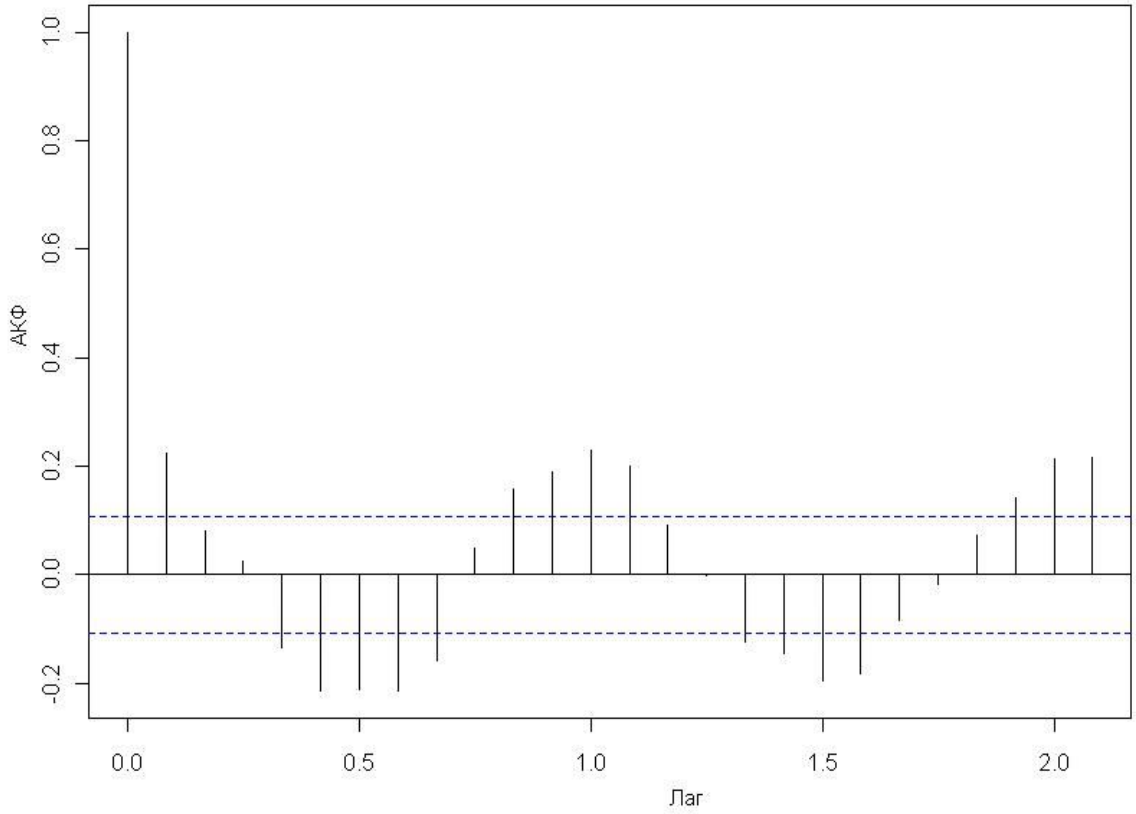


Рис. 7.7. Автокорреляционная функция ряда ПОВТОР ; пунктирными линиями показаны 95% доверительные интервалы, по оси абсцисс лаг в долях календарного года

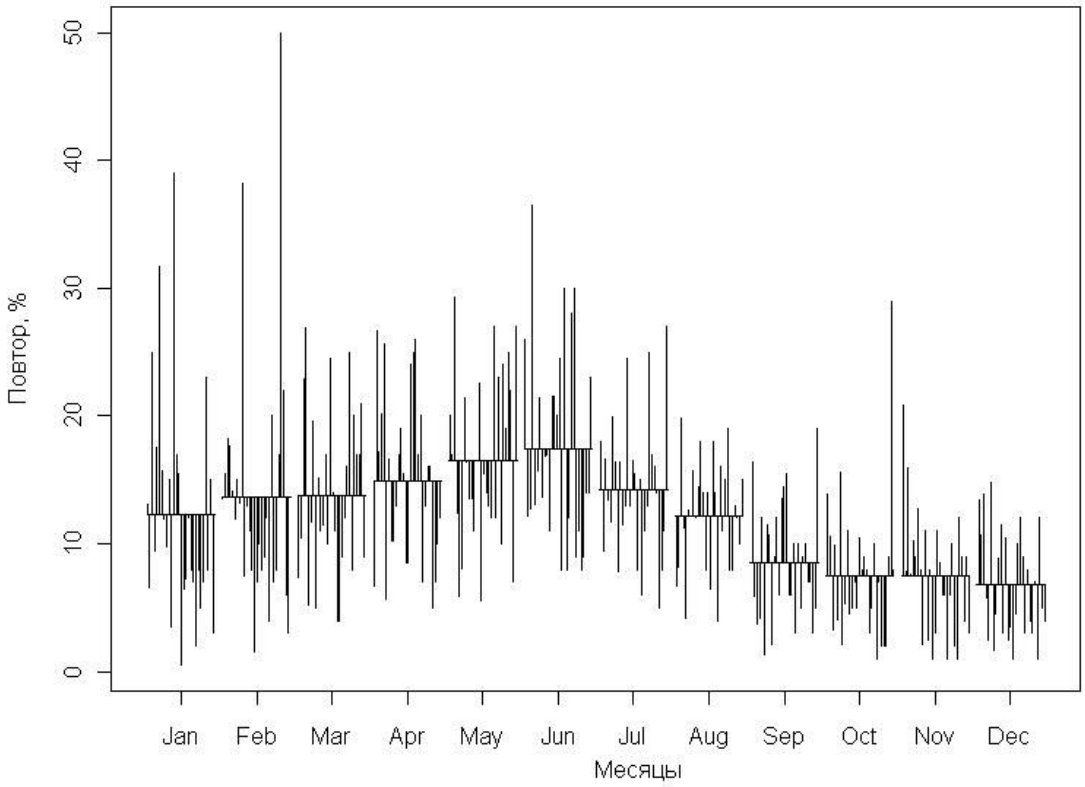


Рис. 7.8. Среднемесячная изменчивость повторяемости северного ветра

Статистическую значимость последовательности коэффициента автокорреляции при различных лагах $\{r_1, \ldots, r_{12}\}_{obs}$ можно проверить с использованием алгоритма рандомизации. Сформируем $B = 1000$ перевыборок, в которых значения исходного временного ряда были случайным образом перемешаны, и рассчитаем матрицу 13×1000 рандомизированных значений $\{r_1, \ldots, r_{12}\}_{ran}$ при справедливости нулевой гипотезы. *P*-значения коэффициентов для каждого лага найдем по обычной формуле $p = (b + 1)/(B + 1)$, где b - число случаев, при которых $\{r_k\}_{ran} \geq \{r_k\}_{obs}$:

Лаг k	1	2	3	4	5	6	7	8	9	10	11	12
$\{r_k\}_{obs}$	0.223	0.082	0.024	-0.13	-0.21	-0.21	-0.21	-0.16	0.047	0.157	0.191	0.228
p_k	**0.001**	0.145	0.670	**0.011**	**0.001**	**0.001**	**0.001**	**0.006**	0.377	**0.007**	**0.001**	**0.001**

Приведенные оценки значимости АКФ в целом соответствуют доверительным интервалам на рис. 7.7, полученным параметрической аппроксимацией статистики Неймана.

При анализе рядов динамики может иметь место взаимное влияние двух процессов при сдвиге последовательностей наблюдений друг относительно друга на некоторый временной промежуток (лаг k), т.е. влияние одного события на другое проявляется с некоторым запаздыванием или опережением. В некоторых случаях наличие временного лага является очевидным: например, в цепи "хищник - жертва" вспышка популяции хищника происходит через определенный промежуток после вспышки популяции жертвы, после чего происходит снижение численности с тем же временным сдвигом.

Взаимная корреляционная функция, или кросс-корреляционная функция (ККФ) определяется для двух стационарных временных рядов как корреляция между x_t и y_{t+k} в зависимости от k. В отличие от автоковариации, кросс-ковариация не является симметричной по k, поэтому ее следует рассматривать как при положительных, так и при отрицательных k. Таблично заданный ряд выборочных коэффициентов ККФ c_k затухает довольно быстро, а наличие пиков указывает на то, что согласованная вариация рядов носит периодический характер. Например, на кросс-спектре рядов СКОРОСТЬ и ПОВТОР (рис. 7.9) можно усмотреть их достоверную взаимосвязь при временном сдвиге через 9 месяцев (хотя это вполне можно счесть как проявление "ложной" корреляции).

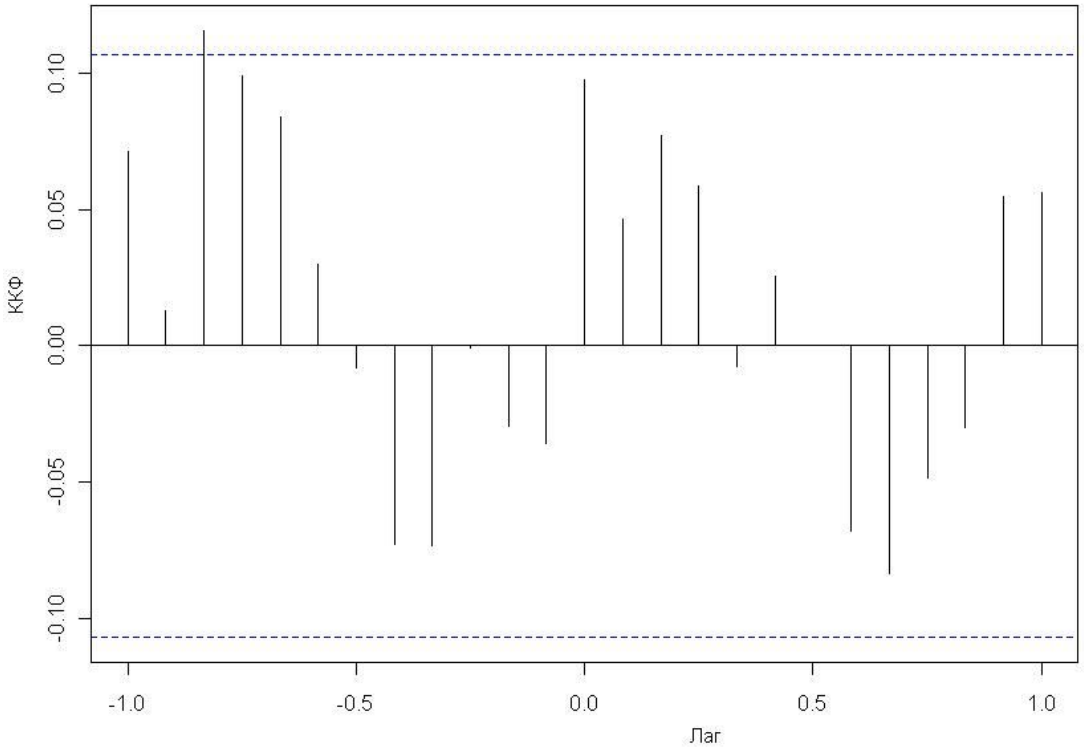


Рис. 7.9. Кросс-корреляционная функция между рядами СКОРОСТЬ и ПОВТОР для северного ветра

Чтобы оценить статистическую значимость АКФ в целом до *m*-го порядка для реального временного ряда, можно было бы применить формулу Бонферрони и установит критическое значение α_m = 0.05/12 = 0.0043 ☺. Однако для этого обычно используют статистику Льюнга-Бокса (Ljung-Box): $Q(r) = n(n+2)\sum_{k=1}^{m} r_k^2/(n-k)$. Если расчетное значение Q-статистики больше 95%-го квантиля распределения χ_m^2 то признается наличие автокорреляции. Ряд ПОВТОР при *m* = 12 является автокоррелированным в целом, т.к. значению статистики $Q(r)$ = 121.1 соответствует весьма малое $p < 0.00001$.

Стационарный процесс с нулевым математическим ожиданием и отсутствием автокорреляции называют "белым шумом". При разработке моделей временных рядов важна необходимость отличать стационарный процесс от нестационарного на основе формальных критериев. Наиболее распространен тест Дики-Фуллера (Dickey, Fuller, 1979), который предполагает, что нестационарность определяется двумя процессами:

$x_t = \mu_0 + \varphi x_{t-1} + \varepsilon_t$ или (7.2)

$x_t = \mu_0 + \mu_1 t + \varphi x_{t-1} + \varepsilon_t$, (7.3)

где ε_t - стационарный ряд остатков ("белый шум"), μ_0 , μ_1 и φ - параметры моделей.

Модель (7.2) описывает марковский процесс или авторегрессию первого порядка. При $|\varphi| < 1$ мы имеем затухающий стационарный процесс, а при $|\varphi| > 1$ – "взрывной" процесс, в котором влияние прошлых флуктуаций усиливается со временем. Авторегрессионный процесс при $\varphi = 1$ и $\mu_0 = 0$ представляет собой случайное блуждание со стохастическим трендом, а при $\mu_0 = 0$ – случайное блуждание с дрейфом. Модель (7.3) является комбинацией линейного тренда и случайного блуждания с дрейфом.

Нулевая гипотеза в тесте Дики–Фуллера состоит в том, что ряд нестационарен и имеет один единичный корень H_0: $\varphi = 1$, $\mu_1 = 0$, а альтернатива предполагает, что ряд стационарен и $|\varphi| < 1$. Для проверки H_0 оценивают критерий $t = (\varphi - 1)/s_\varphi$, сравнивая его с табличными критическими значениями, полученными методом Монте-Карло.

В рассматриваемых примерах оценивалось уравнение регрессии (7.3) и при максимальном лаге *k* = 12 рассчитывались значения статистики Дики-Фуллера *t* и соответствующие ей уровни значимости *p*, на основе чего сделаны следующие выводы:

- для рядов ПОВТОР (t = -4.5, p = 0.01) и РАСХОД ((t = -3.76, p = 0.02) нулевая гипотеза о нестационарности отвергается в пользу альтернативы и процессы считаются стационарными относительно линейного тренда при $\varphi < 1$;
- для ряда СКОРОСТЬ (t = -1.8, p = 0.66) нет оснований отвергнуть H_0 и предполагается нестационарный процесс случайного блуждания с дрейфом ($\varphi = 1$).

Одной из самых интересных процедур поиска закономерностей во временном ряду является выделение периодичностей, толкование смысла которых порождает самые интригующие космогонические предположения. Русский математик Е.Е.Слуцкий (1927), положивший начало анализу периодичностей, писал: «Наличие синусоидальных волн различных порядков, начиная с длинных, обнимающих десятилетия, продолжая циклами примерно от пяти до десяти лет длиною и кончая совсем короткими волнами, остается как факт, требующий объяснения». Например, одиннадцатилетние метеорологические циклы предполагают связь с числом пятен на Солнце, а циклы скачков биологической эволюции в 26 миллионов лет предполагают возможность, что у Солнца есть сопутствующая звезда.

Целью *спектрального* анализа является разложение дисперсии временного ряда σ^2 по различным частотам. При наличии статистически значимых циклических составляющих в соответствующих частотах появляются пики, позволяющие судить о вкладе этих гармоник в дисперсию процесса. Так на графике рис. 7.10а отчетливо видна 12-месячная сезонная периодичность расхода воды на плотине Куйбышевской ГЭС, связанная с весенним паводком. Если имеет место непериодический тренд (или тренд с бесконечным периодом), это обычно сопровождается характерным скачком на нулевой

частоте, т.е. в начале координат спектральной функции, что и показывает, например, спектрограмма ряда СКОРОСТЬ.

Используя особенность тригонометрических функций, которые в определенном диапазоне частот обладают свойством ортогональности, временной ряд $\{x_t\}$, $t = 1,2,\ldots, n$, можно представить как конечный ряд Фурье (Manly, 2007):

$$x_t = A_0 + \sum_{k=1}^{m-1}\{A_k \cos(w_k t) + B_k \sin(w_k t)\} + A_m \cos(w_m t),$$

где $m = n/2$, $w_k = 2\pi k/n$ – все возможные круговые частоты временного ряда. Если, например, $n = 100$ дней, то частота w_1 определяет 100-дневный цикл, w_2 – 50-дневный и т.д. до w_m, которая связна с 2-дневной периодичностью. Здесь мы в правой части имеем n неизвестных коэффициентов, которые вычисляются как

$$A_0 = \bar{x}; \quad A_k = (2/n)\sum_{i=1}^{n} x_i \cos(w_k t); \quad B_k = (2/n)\sum_{t=1}^{n} x_i \sin(w_k t); \quad A_m = (1/n)\sum_{t=1}^{n} x_t (-1)^t .$$

Если представить $S_k^2 = A_k^2 + B_k^2$, то $n\{\sum_{k=1}^{m-1} S_k^2/2 + A_m^2\} = \sum_{t=1}^{n}(x_t - \bar{x})^2$, и мы имеем разложение общей суммы квадратов отклонений членов ряда от его среднего на (m - 1) составляющих изменчивости, объясняемых каждой частотой периодичности.

График зависимости nS_k^2 от частоты Фурье w_k или длины периода называется *периодограммой* и может быть использован для обнаружения и оценки амплитуды гармонической компоненты неизвестной частоты, скрытой в шуме. В определении периодограммы принципиальным является то, что частоты изменяются дискретно, причем наиболее высокая частота составляет 0.5 цикла за весь временной период. Ранее, вводя понятие спектра, мы позволяли частоте меняться непрерывно в некотором диапазоне.

Если график периодограммы слишком "зазубрен" и гармоническая составляющая перераспределена по многим частотам, то визуально оценить наличие циклических процессов трудно. Для этого используют статистики, с помощью которых проверяют нулевую гипотезу стохастичности против альтернативы, что есть, по крайней мере, один периодический компонент. В частности, g-критерий Фишера соотносит максимальную ординату периодограммы к сумме всех пиков: $g = \max_{1<k<m} I(w_k) / \sum_{k=1}^{m} I(w_k)$ и мы можем аналитически найти p-значение вычисленной статистики с использованием нормального распределения.

Более полный и мощный тест на отсутствие периодичности предложен М.Бартлеттом (Manly, 2007) с использованием статистики Колмогорова-Смирнова:

$D = \max(D^+, D^-)$; $\quad D^+ = \max\{j/(m-1) - u_j\}$; $\quad D^- = \max\{u_j - (j-1)/(m-1)\}$,

где $u_j = \sum_{k=1}^{j} S_k^2 / \sum_{k=1}^{m-1} S_k^2$. По существу здесь D^+ – максимальное отклонение ряда u-значений ниже ожидаемого уровня, D^- – то же, но в сторону превышения, а D – полное максимальное отклонение. Разумеется, наиболее корректные оценки p-значений по этим статистикам можно получить с использованием обычного алгоритма рандомизации.

Для рассматриваемого ряда ПОВТОР, периодограмма которого приведена на рис. 7.10б, по обоим критериям установлено существование периодической составляющей: $p < 0.00001$ с использованием g-критерия Фишера в классическом и робастном вариантах и D = 0.2067 при уровне значимости $p = 0.002$, найденном после 500 итераций рандомизационного теста.

Если H_0 отклонена и предполагается наличие периодичности, то необходимо найти частоты, определяющие статистически значимые циклы. Для этого введем величины

$$P_k = S_k^2 / 2\sum_{i=1}^{n}(x_i - \bar{x})^2; \quad \sum_{k=1}^{m} P_k = 1,$$

которые пропорциональны долям общей вариации ряда, связанным с различными частотами. Высокие значения P_k указывают на частоты, определяющие периодичность процесса. Уровни значимости могут быть определены, сравнивая каждую P_k с распределением, полученным для этой статистики после многократного перемешивания порядка следования значений временного ряда (Manly, 2007).

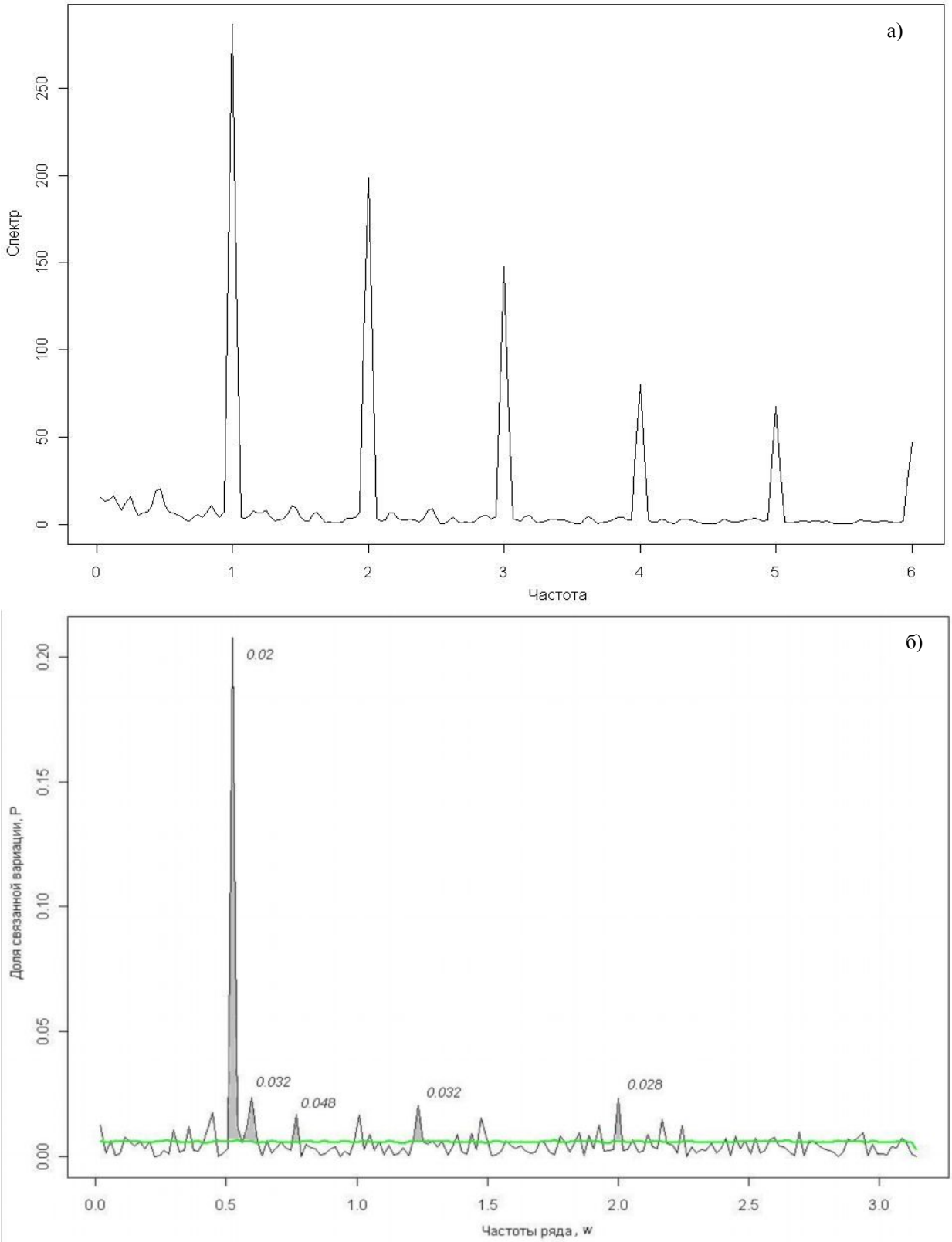


Рис. 7.10. Спектр ряда РАСХОД (а) и периодограмма ряда ПОВТОР (б); зеленой линией показана периодограмма ряда при справедливости нулевой гипотезы, приведены *p*-значения для статистически значимых частот

Для ряда ПОВТОР (см. рис. 7.10б), включающего n= 336 наблюдений, с использованием рандомизационного теста было проанализировано $m = n/2 = 168$ частот. При этом, кроме основной частоты $w_{12} = 0.523$, соответствующей сезонному циклу в 12 месяцев, было выделено еще 4 частоты, которые могут быть связаны с этим процессом на уровне значимости $p < 0.05$ и имеющие более краткие периоды 10.5, 8.2, 5.1 и 3.14

месяцев. И опять уместно задасться вопросом, применимо ли здесь формула Бонферрони, которая предполагает, что если при уровне значимости $\alpha = 0.05$ объявить периодичность на любой частоте случайностью, то пороговая значимость для индивидуальных испытаний на одной частоте составляет в этом примере $\alpha_m = 0.05/168 = 0.00029$ ☺?

К разделу 7.2:

```
load(file="time_ser.RData")
#  Оценка автокорреляции и тест на ее статистическую значимость
acf(WIND.repl) ;   Box.test(WIND.repl, lag=12, type="Ljung-Box")
ccf(WIND.speed,WIND.repl,lag.max=12) #  Оценка кросс-корреляции двух рядов
#  Проверка АКФ(13) с использованием рандомизационного теста
lag=14 ; ac_emp <- acf(WIND.repl,plot = F,lag.max=lag-1)$acf ; N_rand=1000
ac_rand = matrix(rep(0,N_rand*lag), ncol=lag)  ; for (i in 1:N_rand) {
ac_rand[i,] <- acf(ts(sample(WIND$Повтор), frequency=12, start=c(1961,1)),
                                                    plot = F,lag.max=lag-1)$acf }
p <- rep(0,lag) ; for (j in 1:lag) {
             p[j] <- (sum(abs(ac_rand[,j])- abs(ac_emp[j]) >= 0)+1)/(N_rand+1) }
print(cbind(ac_emp, p),3)
library(forecast) ; monthplot(WIND.repl,ylab="Повтор, %",xlab="Месяцы",xaxt="n", type =  "h")
axis(1,at=1:12,labels=month.abb,cex=0.8)  #  График месячной изменчивости
#           Тест Дики-Фуллера на стационарность процесса (сравнение двух вариантов функций)
library(forecast) ; WIND.repl.DF <- adf.test(WIND.repl, alternative = "stationary")
source("DF.r")  ; adf.test.1(WIND.repl,kind=3, k=12)
adf.test.1(WIND.speed,kind=3, k=12) ; adf.test.1(FLOW,kind=3, k=12)
#           Построение спектров и выявление периодичности
spec.pgram(FLOW, spans=3, log="no") ; spec.pgram(WIND.speed, spans=3, log="no")
spec.pgram(WIND.repl, spans=3, log="no")
# Тест по g-критерию Фишера в классическом и робастном вариантах
library(GeneCycle) ; t=1:length(WIND$Повтор) ; fisher.g.test(WIND.repl)
y = robust.spectrum(x=WIND.repl,t=t, algorithm="regression")
pvals = robust.g.test(y = y, perm=TRUE, x=as.matrix(WIND.repl), noOfPermutations = 50,
                    algorithm = "regression", t=t)
#     Рандомизационный тест Манли для анализа периодичности
Cycle.Per <- function (X ,N=length(X), NRAND=100) {
# Возвращает таблицу: W  -  вектор частот, CYCLE - вектор длин периодичностей для P
# P и P0 - ординаты периодограммы для наблюдений и полученных рандомизацией
# SIGP - уровни значимости (% рандомизированных значений, превышающих наблюдаемые)
#  SKS и SIGKS - статистика Колмогорова-Смирнова и уровень ее значимости
   NS = N%/%2 ; N = 2*NS ; XS <- scale(X) ; PLE <- Period(XS[1:N],N)
   SIGKS=1 ; SIGP <- rep(1,NS) ; P0 <- rep(0,NS) ; W = (1:NS)*2*pi/N ; CYCLE = N/(1:NS)
   for (ira in 2:NRAND) { PLR <- Period(sample(XS, N), N)
      if (PLR$SKS > PLE$SKS) SIGKS = SIGKS + 1
      SIGP[which(PLR$P > PLE$P)] + 1 -> SIGP[which(PLR$P > PLE$P)]; P0 <- P0 + PLR$P }
   tab.per <- data.frame(W=W,CYCLE=CYCLE,P=PLE$P,P0=P0/NRAND,SIGP=SIGP,
                       SIGPP=SIGP/NRAND)
   return(list(Period=tab.per,SKS=PLE$SKS,SIGKS=SIGKS/NRAND)) }
cos.per <- function (N, M, i, k){return(ifelse(k==M, (-1)^i/N, cos(2*pi*k*i/N)/M))}
sin.per <- function (N, M, i, k){return(ifelse(k==M, 0, sin(2*pi*k*i/N)/M))}
Period  <- function (XS ,N) {
# Возвращает: P  - вектор ординат периодограммы, пропорциональных объясняемой дисперсии
# SKS   - статистика Колмогорова-Смирнова
       M=N/2 ; S2 <- P <- A <- B <- rep(0,M)
       for ( i in 1:N) {
           for ( k in 1:M) {
                A[k] + XS[i]*cos.per(N, M, i, k)  -> A[k]
                B[k] + XS[i]*sin.per(N, M, i, k)  -> B[k] } }
   S2 <- A*A+B*B; P = S2/2; P[M]=A[M]*A[M] ; U <- cumsum(S2[1:M-1]/sum(S2[1:M-1])); DMAX <- 0
   for ( k in 1:M-1) DMAX <- max(c(DMAX, k/(M-1) - U[k], U[k] - (k-1)/(M-1)))
   return(list(P=P,SKS=DMAX)) }
#  Выполнение расчетов и построение графика периодограммы
Per.Repl <- Cycle.Per(WIND$Повтор, NRAND=500)
plot (Per.Repl$Period[,1], Per.Repl$Period[,3],"l")
```

```
lines(Per.Repl$Period[,1], Per.Repl$Period[,4], lwd=2, col="green")
```

7.3. Модели временных рядов: бутстреп и прогнозирование

При построении моделей аппроксимации и прогнозирования необходимо учитывать наличие автокорреляции временных рядов и периодичность изменения анализируемой случайной величины. Рассмотрим три типа таких моделей: а) линейную регрессию с детерминированными компонентами, б) адаптивные сезонные модели, основанные на экспоненциальном сглаживании и в) комбинированную модель авторегрессии - интегрированного скользящего среднего.

Ранее (рис. 7.10а) было отмечено, что временной ряд РАСХОД имеет отчетливую 12-месячную периодичность, связанную с массовым сбросом вод Куйбышевского водохранилища во время весеннего паводка. Простейшие модели цикличности могут быть основаны на фиктивной переменной $z_t = t/12$, $t = 1,2, \ldots, n$ и иметь следующий вид:

- без учета непериодической тенденции $x = \beta_0 + \beta_1\sin(2\pi z_t) + \beta_2\cos(2\pi z_t) + \varepsilon$;
- с учетом линейного тренда $x = \beta_0 + \beta_1\sin(2\pi z_t) + \beta_2\cos(2\pi z_t) + \beta_3 t + \varepsilon$.

Подбор коэффициентов линейной модели с трендом для этого ряда дает следующие результаты:

Факторы	Коэффициенты	Ошибка S_{bj}	t-критерий	p-значение
Св. член β_0	19.1	1.09	17.6	< **0.00001**
$\sin(2\pi z_t)$	4.453	0.769	5.78	< **0.00001**
$\cos(2\pi z_t)$	-7.812	0.769	-10.15	< **0.00001**
t	0.0051	0.0049	1.051	0.294

при приведенном коэффициенте детерминации $R^2 = 0.259$ и критерии Акаике AIC = 2913.

Разумеется, в этих условиях следует отдать предпочтение модели без учета тренда (т.е при $\beta_3 = 0$), сумма квадратов отклонений которой статистически значимо не отличается от модели с трендом ($F = 1.1$, $p = 0.29$), но имеющей меньший AIC = 2912. График аппроксимации ряда РАСХОД этой моделью представлен на рис. 7.11.

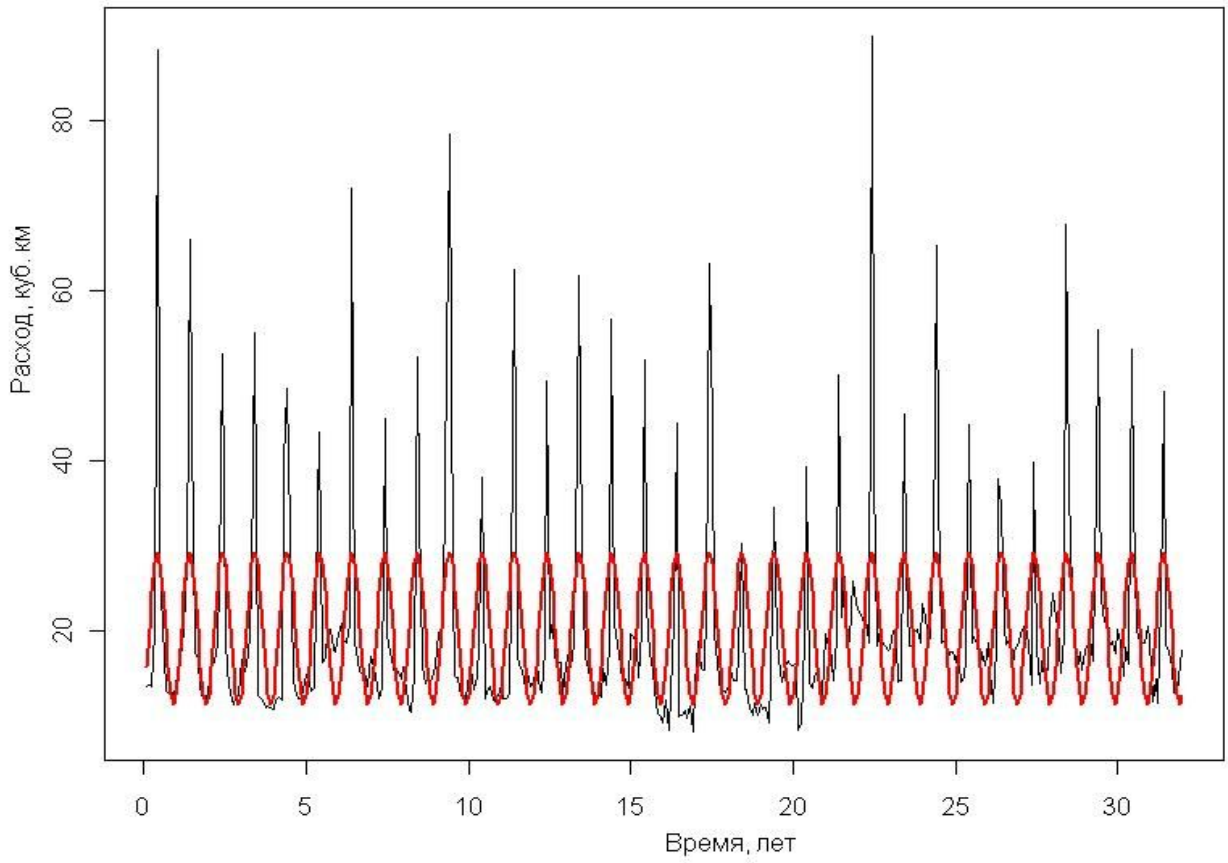


Рис. 7.11. Аппроксимация ряда РАСХОД периодической функцией с фиксированными факторами

Общий смысл использования процедуры бутстрепа для анализа моделей временных рядов остается без изменений: необходимо многократно извлекать псевдовыборки из распределения данных, насколько возможно близкого к эмпирической последовательности, и повторять расчет оцениваемых параметров для каждой такой повторности испытаний. Однако, поскольку исходный временной ряд не является независимым, то мы не можем получить корректную его имитацию простым случайным выбором с возвращением. Необходимо использовать более сложные алгоритмы генерации псевдовыборок, которые бы не разрушали автокорреляционную или периодическую зависимость, существующую в эмпирическом ряду.

Наиболее близок к классической непараметрической процедуре простой блочный бутстреп. Если для ряда наблюдается значимая периодичность длиной l, то мы можем разделить исходный ряд на последовательность из $m = n/l$ неперекрывающихся блоков. Каждая новая псевдоповторность исходного ряда $\tilde{x}_l^n$ конструируется из этих блоков с использованием внутри их стандартного алгоритма случайного выбора с возвращением. Отметим, что тогда между блоками в некотором смысле сохраняются эмпирические зависимости в периодичности наблюдений. Для более сложных случаев автокорреляции и периодичности разрабатываются варианты бутстрепа с перекрывающимися блоками и лаговым сдвигом, циркуляционного бутстрепа и т.д.

С использованием простого блочного бутстрепа можно выполнить оценку доверительных интервалов параметров β_1, β_2 представленной выше регрессионной модели временного ряда РАСХОД без учета линейного тренда – см. рис. 7.12. Поскольку доверительные границы коэффициентов располагаются вдалеке от нулевой зоны, то это служит решающим аргументом в пользу статистической значимости уравнения.

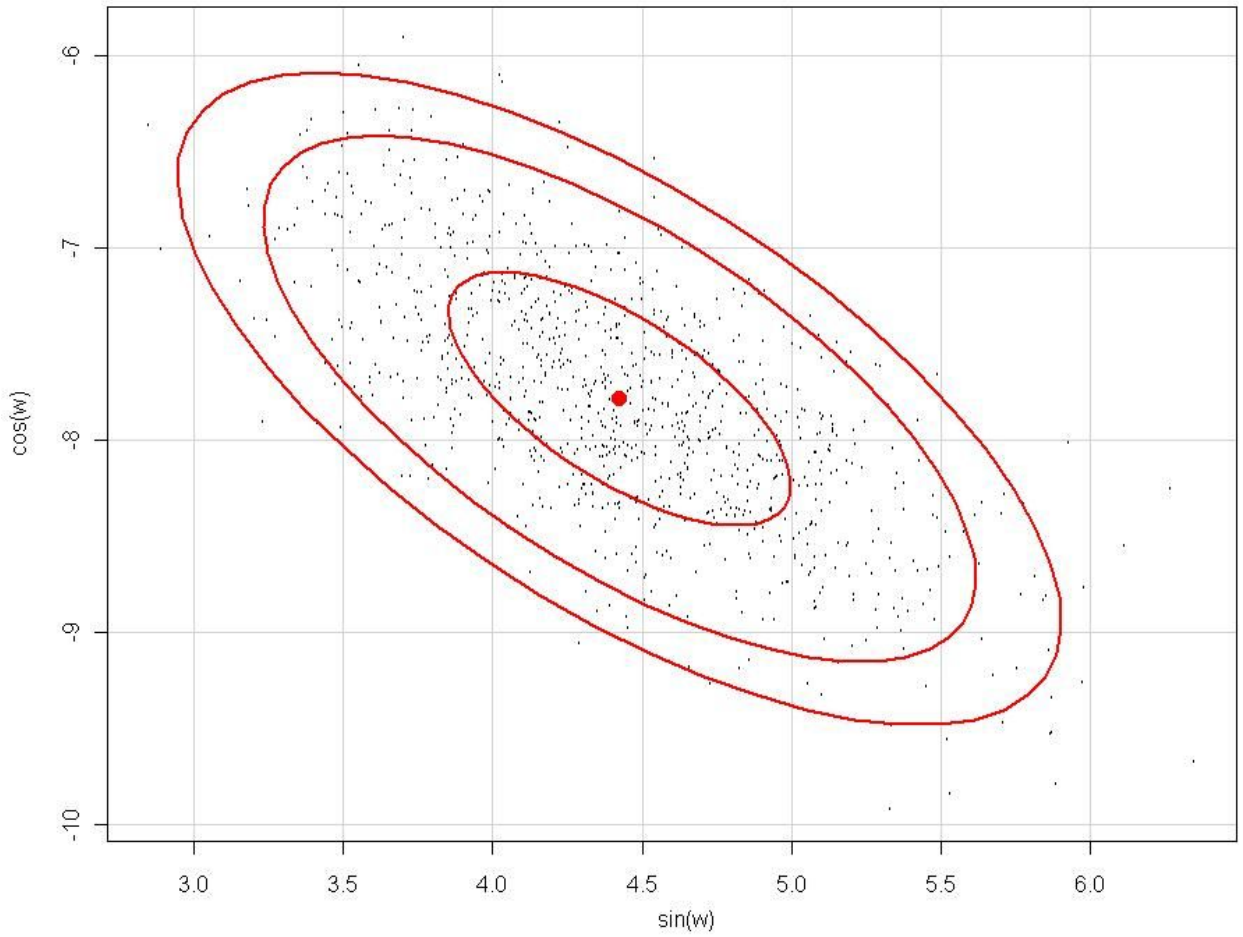


Рис. 7.12. Диаграмма рассеяния бутстреп-оценок коэффициентов регрессии периодического тренда ряда РАСХОД и эллипсы доверительных интервалов с вероятностями 0.5, 0.95 и 0.99

Предположим теперь, что временной ряд описан произвольной моделью с предикторами, являющимися функциями t, и были получены оценки параметров на основе выборочных данных $\mathbf{x} = (x_1, ..., x_n)$. Это позволяет построить прогноз на будущее, например на период $n + h$. Ошибку прогноза можно представить как сумму двух отдельных ошибок: (а) ошибки, связанной с тем, что истинные параметры β модели неизвестны и вместо них используются оценки b (т.е. ошибка, обусловленная выборкой) и (б) будущей ошибки прогнозируемого наблюдения, обусловленной изменениями характера моделируемого случайного процесса на интервале $n \div h$, которые не могли быть учтены моделью.

При конструировании моделей с целью минимизировать ошибку регрессии вступают в конкуренцию два принципа: желание учесть как можно полнее принципиальную сложность реальной системы и принцип простоты. Принцип простоты или экономии предусматривает, что из двух или нескольких практически эквивалентных хороших моделей для расчетов прогнозов следует выбирать простейшую (в англоязычной литературе этот принцип известен как Principle of Parsimony, что не имеет отношение к красивой итальянской фамилии Парсимони). Несовместимость "простоты" модели и точности решения задачи проявляется в высказывании академика А.А.Самарского (цит. по Розенберг и др., 1994): «... исследователь постоянно находится между Сциллой усложненности и Харибдой недостоверности. С одной стороны, построенная им модель должна быть простой в математическом отношении, чтобы ее можно было исследовать имеющимися средствами. С другой стороны, в результате всех упрощений она не должна утратить и ‘рациональное зерно’, существо проблемы.».

Пусть на основе временного ряда x_t ($t = 1, \ldots, n$) оцениваются наиболее вероятные прогнозируемые значения в произвольных точках h будущего периода $\hat{x}_{t+h|t}$. Некоторые методы прогноза очень просты и при этом иногда могут быть удивительно эффективными. Следующие три простейших метода мы будем использовать как точки отсчета для других, более продвинутых моделей:

1. *Метод средних*: прогноз равен среднему значению ряда $\hat{x}_{t+h|t} = \bar{x}$ (т.е. считается, что развитие процесса совершенно стабилизировалось и на любой период будущего прогнозируются одни и те же значения отклика).
2. *"Наивный" метод с дрейфом*: к последнему наблюдаемому значению ряда добавляется среднее приращение Δx_t, зависящее от основной тенденции в историческом периоде $$\hat{x}_{t+h|t} = x_n + \frac{h}{n-1}\sum_{i=2}^{n} x_i - x_{i-1} = x_n + h(x_n - x_1)/n$$
3. *Сезонный "наивный" метод*: прогнозом считаются значения, зафиксированные в те же самые календарные даты последнего наблюдаемого цикла $\hat{x}_{t+h|t} = x_{n+h-km}$, где m – период сезонности (например, $m = 12$ месяцам в году), $k = (h - 1)/m + 1$.

Как отмечалось выше, одной из альтернатив по отношению к методам прогнозирования на основе детерминированных моделей является использование различных процедур аппроксимации. Поскольку многие процессы содержат тенденцию, сочетающуюся с ярко выраженными сезонными колебаниями, эффективным способом их описания являются адаптивные сезонные модели, основанные на экспоненциальном сглаживании. Особенность адаптивных моделей заключается в том, что по мере поступления новой информации происходит "мягкая" корректировка параметров модели, их приспособление, адаптация к изменяющимся во времени условиям развития процесса.

Базовая модель с аддитивным сезонным эффектом, предложенная Г. Тейлом и С. Вейджем (Тейл, 1971), имела вид : $$x_t = f_t + g_t + \varepsilon_t,$$
где f_t отражает тенденцию развития процесса, g_t, g_{t-1}, ..., g_{t-k+1} – аддитивные коэффициенты сезонности; k – количество опорных временных интервалов (фаз) в полном сезонном цикле; ε_t – "белый шум".

С. Хольт и его студент П. Винтерс (Holt, 1957; Winters, 1960; цит. по Hyndman et al., 2008) развили эти идеи и разработали весьма изощренный метод, позволяющий успешно справляться с среднесрочными и с долгосрочными прогнозами. Модели Хольта-Винтерса для любой точки ряда последовательно вычисляют сглаженное значение прогнозируемой величины на основе данных предыдущего периода с учетом тренда и сезонных изменений. Наиболее полная модель включает уравнение для прогнозирования и три уравнения для расчета параметров сглаживания: α – для оценки средневзвешенного уровня, β – для оценки смещения за счет тренда и γ – для оценки сезонной компоненты.

Существует две модификации моделей, отличающиеся по характеру сезонных изменений. Аддитивный метод предпочтителен, если сезонный фактор действует примерно постоянно на протяжении всего ряда, в то время как мультипликативная модель используется, когда интенсивность сезонных изменений пропорциональна величине x_t.

Аддитивная модель А имеет вид: $\hat{x}_{t+h\,|\,t} = l_t + hb_t + s_{t-m+h_m^+}$,

где $l_t = \alpha(x_t - s_{t-m}) + (1-\alpha)(l_{t-1} + b_{t-1})$; $b_t = \beta(l_t - l_{t-1}) + (1-\beta)b_{t-1}$; $s_t = \gamma(x_t - l_{t-1} - b_{t-1}) + (1-\gamma)s_{t-m}$.

Сезонные индексы $s_{t-m+h_m^+}$ рассчитываются по точкам ряда $h_m^+ = (h-1)/m+1$.

Мультипликативная модель М записывается в несколько иной форме:

$$\hat{x}_{t+h\,|\,t} = (l_t + hb_t)s_{t-m+h_m^+};\ \ s_t = \gamma\, x_t/(l_{t-1} - b_{t-1}) + (1-\gamma)s_{t-m},$$

а остальные компоненты рассчитываются идентично аддитивному методу.

Существует более 15 версий сокращенных моделей (Hyndman et al., 2008), в которых не учитывается та или иная составляющая (т.е. коэффициенты α, β или γ приравниваются 0), например:

- M(N, N) – простое экспоненциальное сглаживание без тренда и сезонности (β =0; γ=0);
- M(N, M) – мультипликативная сезонная модель без тренда (β = 0);
- M(A, N) – аддитивная модель тренда без учета сезонного фактора (β = 0) и т.д.

Для тестирования эффективности различных прогнозирующих моделей разделим ряд ПОВТОР на две части: по 306 точкам будет выполняться обучение моделей, а 30 точек с июля 1986 по декабрь 1988 будут выделены для экзамена.

Полная аддитивная модель Хольта-Винтерса была получена с коэффициентами α = 0.00995, β = 0.209, γ = 0.217 и соответствующим набором сезонных индексов. График наблюдаемых и прогнозируемых значений для периодов обучения и экзамена представлен на рис.7.13а. Для сравнения там же представлены графики прогнозируемых значений в экзаменуемом периоде для трех "простейших" алгоритмов прогнозирования (б) и для мультипликативной модели Хольта-Винтерса (в).

С помощью функции ets(…) пакета forecast может быть осуществлен автоматический перебор возможных значений коэффициентов α, β и γ при различных лагах и идентификация наилучшей модели Хольта-Винтерса с точки зрения минимума ошибки сглаживания при обучении (Hyndman et al., 2008). В нашем примере наименьшую величину AIC-критерия доставляет аддитивная модель без учета тренда (β =0).

Наконец, еще одной альтернативой простым параметрическим моделям являются стохастические итеративные модели, в основе которых используются следующие преобразования, трансформирующие последовательность случайных независимых импульсов a_t в традиционно рассматриваемый стационарный процесс:

- фильтр авторегрессии (АР) или марковский процесс, в котором текущее значение ряда x_t выражается в виде конечной линейной совокупности предыдущих значений процесса $x_{t\,-1}$, $x_{t\,-2}$, … плюс импульс a_t;
- фильтр скользящего среднего (СС), в котором процесс x_t образуется из белого шума a_t как взвешенная сумма предыдущей последовательности импульсов a_t, $a_{t\,-1}$, $a_{t\,-2}$ ..

Для описания как стационарных, так и нестационарных рядов со стационарными приращениями d-го порядка и рациональным спектром, наиболее эффективна

комбинированная модель АРИСС (ARIMA) авторегрессии-интегрированного скользящего среднего, разработанная Дж.Боксом и Г.Дженкинсом (1974). Построение и диагностика адекватности таких моделей подробно описана, например, в книге (Шипунов и др., 2012).

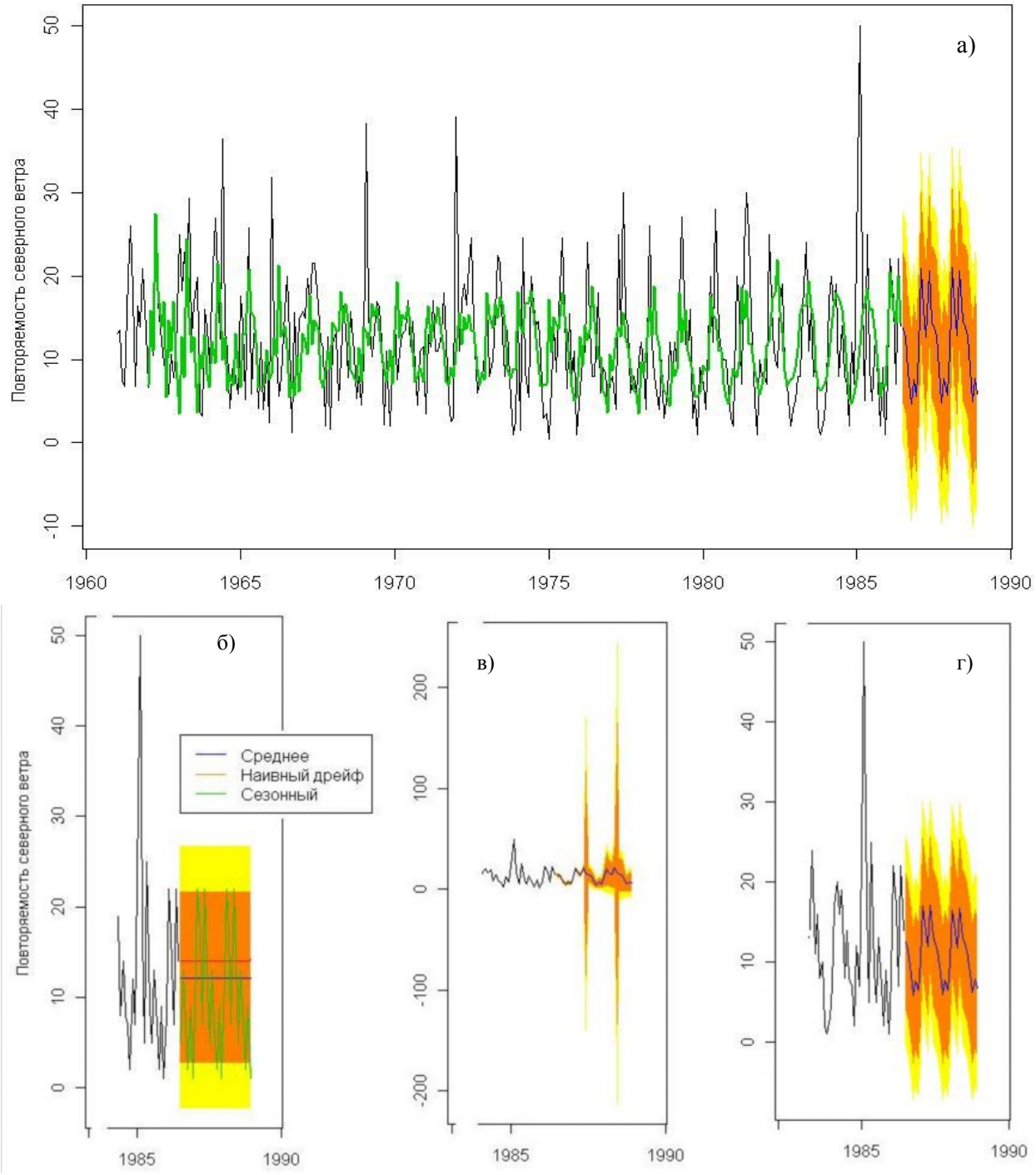


Рис. 7.13. Прогноз значений ряда ПОВТОР, выполненный различными алгоритмами:
а) вверху - аддитивной моделью Хольта-Винтерса для обучаемого (зеленый) и экзаменационного (синий) периодов; внизу - б) с помощью четырех простейших алгоритмов;
в) по мультипликативной модели Хольта-Винтерса; г) по модели ARIMA;
показаны 95% (желтым) и 80% (оранжевым) доверительные интервалы прогнозируемой величины

Схематично несезонную модель Бокса-Дженкинса записывают как ARIMA(p, d, q), где p – порядок авторегрессии, d – порядок разности, q – порядок скользящего среднего. При необходимости учесть периодическую составляющую используют расширенный сезонный вариант ARIMA(p, d, q)(P, D, Q)$_m$, где m –частота, определяющая цикличность. Очевидно, что осуществить вдумчивую идентификацию всех 6 параметров модели

является непростой работой. Воспользуемся однако возможностью, предоставляемую функцией auto.arima(…) пакета forecast, которая автоматически подбирает значения *p, d, q, P, D* и *Q* на основе теста единичного корня и AIC-критерия.

В результате для ряда ПОВТОР при минимуме AIC=2049.2 получаем достаточно экономичную модель ARIMA(1, 0, 0)(2, 0, 2)$_{12}$ с ненулевым средним. Здесь основной моделируемый процесс соответствует марковскому процессу авторегрессии 1-го порядка: $x_t = 0.0378\ x_{t-1} + \varepsilon_t$, а сезонные термы основаны на авторегрессии 2-го порядка и процессе скользящего среднего 2-го порядка. График прогнозируемых значений в экзаменуемом периоде представлен на рис. 7.13г.

Анализ ошибок на обучающей выборке и оценка доверительных интервалов могут сыграть свою определенную роль, но окончательное заключение о точности прогноза всегда основывается на сравнении наблюдаемых x_i и прогнозируемых $\hat{x}_i$ значений с использованием независимой внешней выборки, т.е. вычислении отклонений $e_i = x_i - \hat{x}_i$ для экзаменационной последовательности. При этом различают три группы оценок:

1. Оценки, зависящие от шкалы измерений, из которых традиционными являются среднее абсолютное отклонение $E_{MA}=\overline{|e_i|}$ и корень из среднего квадратичного отклонения $E_{RMS}=\sqrt{\overline{e_i^2}}$.
2. Оценки ошибок, выраженные в процентах, которые не зависят от шкалы измерений и используют $p_i = 100\ e_i / x_i$ %. Эти ошибки нельзя получить, если $x_i \cong 0$. Наиболее широко используется среднее абсолютное отклонение $E_{MAP}=\overline{|p_i|}$.
3. Оценки ошибки в долях шкалы основаны на отношении отклонений прогноза к среднему отклонению анализируемой случайной величины $q_i = e_i(n-1)/\sum_{i=2}^{n}|x_i - x_{i-1}|$. При этом рекомендуется использовать среднее абсолютное шкалированное отклонение $E_{MAS}=\overline{|q_i|}$ или статистику Тейла $U=\{\sum_{i=1}^{n+1}[(\hat{x}_{i+1}-x_{i+1})/x_i]^2 / \sum_{i=1}^{n+1}[(x_{i+1}-x_i)/x_i]^2\}^{0.5}$.

Перечисленные статистики были рассчитаны нами для 30 точек экзаменационной выборки ряда ПОВТОР и сведены в табл. 7.1. Результаты оказались не слишком воодушевляющими для любителей серьезной математики: продвинутые модели Хольта-Винтерса при прогнозе сезонных колебаний часто попадали в диссонанс с реальными наблюдениями, а один из лучших прогнозов соответствовал постоянному значению для всего ряда, близкому к среднему. Зато оправдала надежды модель ARIMA, которая позиционируется иногда как естественное обобщение всех остальных моделей, включая адаптивные процедуры сезонного сглаживания с трендом.

Таблица 7.1. Оценки ошибочного прогноза значений ряда ПОВТОР тремя простейшими алгоритмами и различными моделями для экзаменационной выборки (обозначения по тексту)

Алгоритм прогноза	E_{MA}	E_{RMS}	E_{MAP}	E_{MAS}	U Тейла
Метод средних	6.1	7.5	90.23	0.869	0.94
Наивный метод с дрейфом	6.84	7.92	110.28	0.975	1.066
Сезонный наивный метод	7.33	9.77	94.8	1.044	1.436
Полная аддитивная модель	6.21	8.317	86.35	0.885	1.277
Мультипликативная модель	6.27	8.237	87.30	0.893	1.279
Аддитивная модель без тренда	6.13	8.41	81.05	0.873	1.274
Модель без тренда и сезонности	6.0	7.49	87.06	0.85	0.925
Модель ARIMA	5.83	7.725	75.56	0.83	1.076

К разделу 7.3:

```
load(file="time_ser.RData")
```

```
#     Модель периодического тренда
index <- 1:length(OUT$Расход) ; time=(index)/12
mod.trend <-lm(OUT$Расход~index + sin(time*2*pi)+cos(time*2*pi))
summary(mod.trend) ; AIC(mod.trend)
model<-lm(OUT$Расход~sin(time*2*pi)+cos(time*2*pi))
summary(model); AIC(model) ; anova(mod.trend,model)
plot(time, OUT$Расход, type="l") ; lines(time, predict(model), col="red", lwd=2)
# Простой поблочный бутстреп: на входе - временной ряд и длина блока
rblockboot <- function(ts,block.length,len.out=length(ts)) {
   blocks.needed <- (len.out%/%block.length) ; n <- block.length*blocks.needed
   x <- ts[1:block.length]
   for (l in 2:blocks.needed) { x <- cbind(x, ts[((l-1)*block.length+1):(l*block.length)]) }
   picked.blocks <- sample(1:blocks.needed,replace=TRUE) ;   x.boot <- x[,picked.blocks]
   x.vec <- as.vector(x.boot) ;   return(x.vec)  }
library(car) ; N_Boot = 1000  ; A.boot <- B.boot <- rep(0,N_Boot)
for (i in 1:N_Boot) { boot <- rblockboot(OUT$Расход,12)
       model.boot<-lm(boot~sin(time*2*pi)+cos(time*2*pi))
       A.boot[i] <- model.boot$coefficients[2]; B.boot[i] <- model.boot$coefficients[3]}
data.ellipse(A.boot, B.boot, xlab="sin(w)", ylab="cos(w)",
                    cex=.3, levels=c(.5, .95, .99), robust=T)
#  Три прогноза простейшими алгоритмами
WIND.m1  <- meanf(WIND.tr, h=30) ; plot(WIND.m1, plot.conf=TRUE)
WIND.m2 <- rwf(WIND.tr, h=30, drift=TRUE) ; lines(WIND.m2$mean,col=2)
WIND.m3 <- snaive(WIND.tr, h=30) ;  lines(WIND.m3$mean,col=3)
legend("topleft",lty=1,col=c(1,2,3),   legend=c("Среднее","Наивный дрейф","Сезонный"))
accuracy(WIND.m1, WIND.ex)   #  Оценка ошибки прогноза
accuracy(WIND.m2, WIND.ex) ; accuracy(WIND.m3, WIND.ex)
#  Модели Хольта-Винтерса :           а) без тренда и сезонности
WIND.HW <- HoltWinters(WIND.tr,  beta=FALSE, gamma=FALSE)
WIND.forecast <- forecast.HoltWinters(WIND.HW, h=30) ; plot(WIND.forecast)
accuracy(WIND.forecast, WIND.ex)
WIND.HW <- HoltWinters(WIND.tr)  #   б) полная аддитивная
WIND.forecast <- forecast.HoltWinters(WIND.HW, h=30) ; plot(WIND.forecast)
accuracy(WIND.forecast, WIND.ex)
WIND.HW <- HoltWinters(WIND.tr,  seasonal = "multiplicative") #   в) мультипликативная
WIND.forecast <- forecast.HoltWinters(WIND.HW, h=30) ; plot(WIND.forecast)
accuracy(WIND.forecast, WIND.ex)
ets(y = WIND.tr, damped = FALSE)  #   Подбор лучших моделей
WIND.HW <- HoltWinters(WIND.tr,  beta=FALSE)  #   б) аддитивная без тренда
WIND.forecast <- forecast.HoltWinters(WIND.HW, h=30) ; plot(WIND.forecast)
accuracy(WIND.forecast, WIND.ex)
#  Модель ARIMA
WIND.arima <- auto.arima(WIND.tr) ; WIND.forecast <- forecast(WIND.arima, h=30)
plot(WIND.forecast) ; summary(WIND.forecast); accuracy(WIND.forecast, WIND.ex)
```

7.4. Анализ главных компонент и многомерные временные ряды

Одним из эффективных способов анализа временных рядов для выявления внутренне присущих им закономерностей является его разложение на естественные ортогональные составляющие методом главных компонент (преобразование Карунена-Лоэва, сингулярный спектральный анализ). Этот подход применим к любому временному ряду, не требует его стационарности, как, например, спектральный анализ, автоматически выявляет тренды и позволяет получать многомерные представления временного ряда – фазовые портреты, дающие возможность визуального изучения траектории ряда в многомерном пространстве его состояний (Ефимов и др., 1988; Пузаченко, 2004).

Метод главных компонент, описанный в разделе 6.2 и используемый для редукции размерности динамических систем, заключается в поиске координатных осей, доставляющих максимальную дисперсию при проецировании на них траектории ряда. Максимизация автоковариации вместо дисперсии приводит к методу *гладких* компонент

(Ефимов, Ковалева, 2007), который также может быть весьма полезен при анализе внутренних закономерностей динамики и структуры популяций.

Сущность получения фазовых портретов одномерного ряда заключается в следующем. Если ряд имеет ярко выраженную периодичность колебаний (например, сезонные генерации численности), то в каждый момент времени t можно выделить характерный вектор предыстории процесса (x_t, x_{t-1}, ... , x_{t-l}). Параметр l называется лагом (запаздыванием) и является многомерной характеристикой процесса. Полученные векторы сводятся в таблицу, имеющую n - l строк (объектов) и l + 1 столбцов (признаков):

x_{t+1}	x_t	x_{t-1}	x_{t-2}	…	x_1
x_{t+2}	x_{t+1}	x_t	x_{t-1}	…	x_2
…	…	…	…	…	…
x_{n-1}	x_{n-2}	x_{n-3}	x_{n-4}	…	x_{n-l-1}
x_n	x_{n-1}	x_{n-2}	x_{n-3}	…	x_{n-l}

Если временной ряд порождается некоторой динамической системой с конечным числом параметров, то совокупность отрезков его предысторий можно рассматривать как точки l-мерного фазового пространства. Соединяя их последовательно фрагментами поверхностей или сплайнами, получим траекторию ряда в этом пространстве. Например, в работе (Schaffer, 1984) исследовалась трехмерная траектория (x_t, x_{t-i}, x_{t-2i}) заготовок шкур канадской рыси. Однако использование компьютерной графики в многомерных случаях, как правило, затруднено, поэтому разумно выполнить редукцию лаговой таблицы.

Обработка матрицы лагов методом главных компонент приводит к появлению матрицы счетов **U** тех же размеров. Новые признаки (или компоненты u_{tj}) не коррелируют между собой и являются линейными комбинациями исходных наблюдений ряда:

$$u_{tj} = \sum_{i=0}^{l} p_{ij} x_{t-i}\,, \qquad j = 0, \ldots, l, \qquad t = l + 1, \ldots, n,$$

p_{ij} – нагрузки (см. раздел 6.2).

Вектор первой компоненты имеет максимально возможную из всех остальных линейных комбинаций дисперсию, второй – максимальную дисперсию из линейных комбинаций, ортогональных первой, и так далее.

Так как каждая из полученных компонент является, в свою очередь, новым временным рядом, то ее поведение можно исследовать в зависимости от любой другой компоненты, получая двухмерные фазовые портреты. Кроме того, имеет смысл использовать редуцированные переменные в качестве предикторов, так как за каждой компонентой предположительно стоит порождающая ее самостоятельная и статистически независимая от других причина (Ефимов и др., 1988).

Рассмотрим сопряженный ряд, содержащий результаты экспедиционных наблюдений на ст. 34 Куйбышевского водохранилища и состоящий из четырех показателей: NH_4^+ и Fe – концентрация ионов аммония и железа (мкг/л), NCAL и NROT – значения биомассы каляноид (*Calanoida*) и ротаторий (*Rotatoria*) (г/м3) в пробах воды за каждый из 6 месяцев (май-октябрь) вегетационного периода с 1961 по 1984 гг.; всего 108 точек. Оценим характер сезонной изменчивости этих факторов методом сглаживания локальной регрессией – рис. 7.14. Очевидно, что для биомассы каляноид наблюдается отчетливая цикличность с максимум популяционной плотности в августе. Аналогично максимум содержания ионов железа приходится на майский паводок, а численность ротаторий определяется как весенней, так и осенней генерацией.

Выполним построение матрицы предысторий с лагом l = 5 и последующим выделением главных компонент. Известно, что разложение на главные компоненты при возрастании n и l сходится к ряду Фурье, а собственные векторы стремятся к отрезкам синусоид. Вследствие ортогональности синуса и косинуса каждой частоте отвечают две сопряженные компоненты. Если взять в качестве базовых первые две главных компоненты лаговой матрицы ряда NCAL, объясняющие 58% дисперсии ряда, то на сформированном фазовом портрете (рис. 7.15) прослеживается отчетливая сезонная закономерность в виде 6-точечной спирали. Эта цикличность в наибольшей мере характерна до середины 70-х годов (точки 6-74), а в последующий период спираль

начинает "скручиваться в клубок", что можно объяснить нарушением естественных механизмов развития планктоценоза после строительства каскада ГЭС на Волге и Каме.

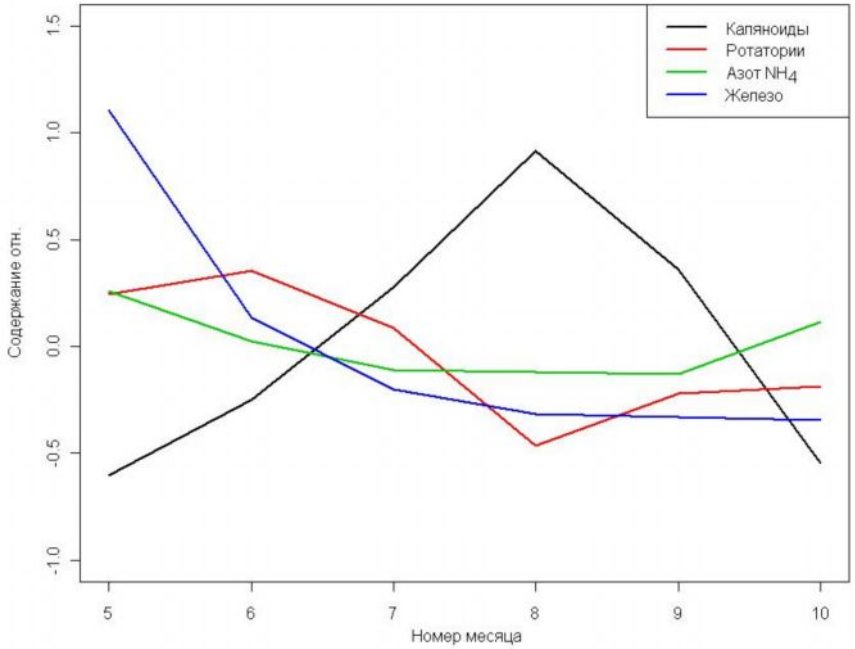


Рис. 7.14. Сезонная динамика изменчивости четырех гидробиологических и гидрохимических показателей на ст. 34 Куйбышевского водохранилища (1958-1984 гг)

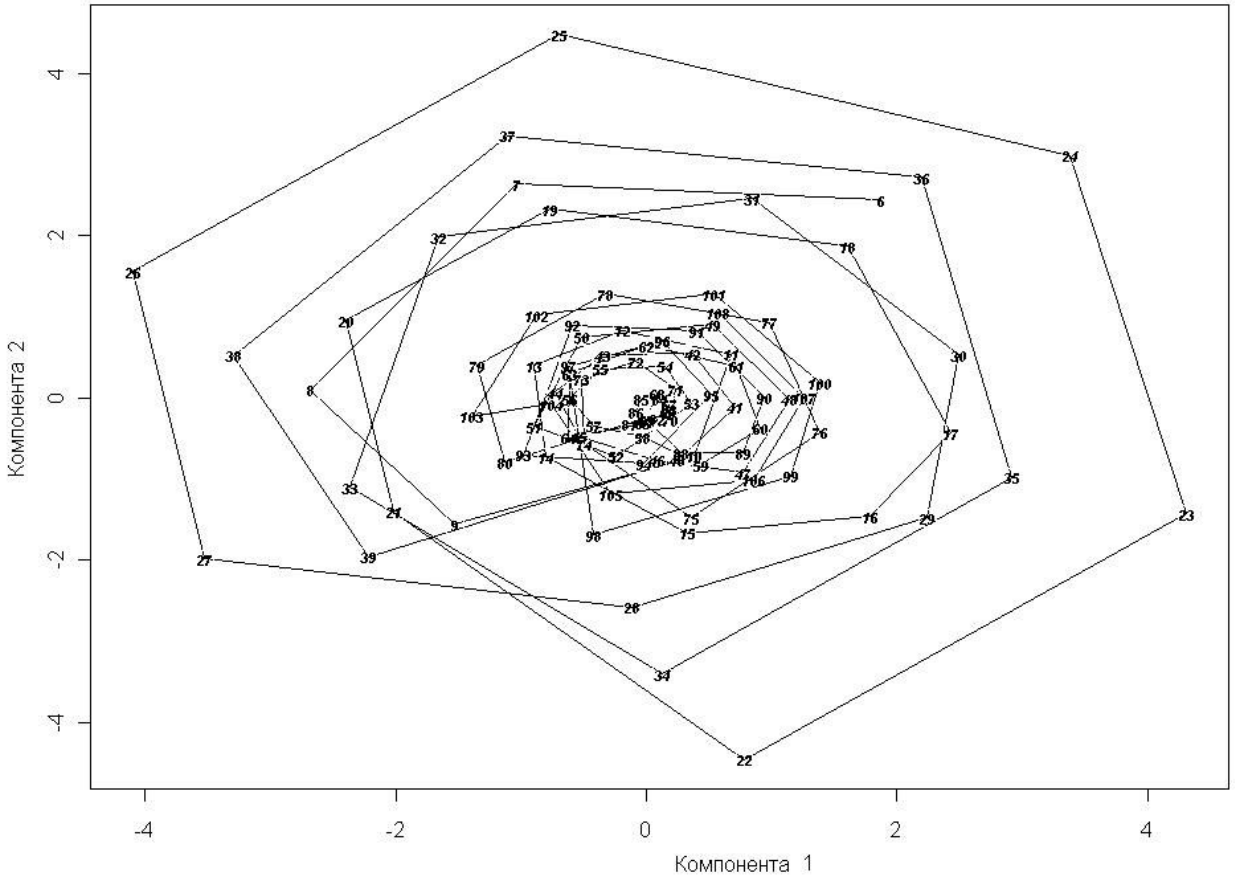


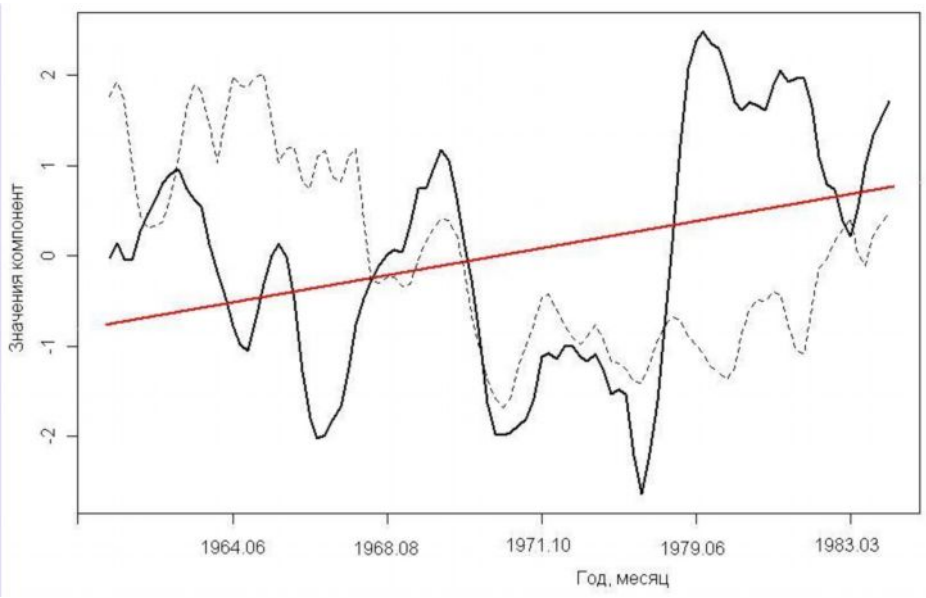
Рис. 7.15. Фазовый портрет динамики численности каляноид в пространстве первых двух главных компонент (точка 6 – май 1962 г., точка 108 – октябрь 1984, цикл наблюдений – 6 месяцев)

Аналогичное разложение динамики численности ротаторий по ортогональному базису главных компонент приводит к фазовой траектории в виде хаотичных зигзагообразных перемещений. Здесь, правда, следует отметить две важнейшие проблемы,

не решенные к настоящему времени: как выбрать величину лага *l* в неочевидных случаях и как количественно оценить статистическую значимость отличия траектории с визуально выраженной закономерностью от хаотичного фазового портрета.

Методом главных компонент можно обрабатывать и совокупности взаимосвязанных временных рядов. В этом случае информация представляется в виде матрицы, в которой объектами являются отсчеты (например, месяцы), а признаками служат исследуемые временные ряды. После расчета собственных значений и векторов корреляционной матрицы бóльшая часть информации оказывается сосредоточенной в первых компонентах и любую из них можно анализировать как новый временной ряд.

Выполним оценку совместную тенденцию динамики четырех показателей сопряженного ряда за наблюдаемый период (1961-1984 гг.). Предварительно выделим тренд каждого фактора, исключив с помощью функции decompose() периодическую составляющую и случайные флуктуации. Первая главная компонента (рис. 7.16) обобщает изменчивость концентрации загрязняющих веществ и (с обратным знаком) численности зоопланктонных сообществ. В целом ее можно интерпретировать термином "общая экологическая напряженность", которая имеет выраженную тенденцию к росту. Наряду с этим легко выделить отдельные периоды, связанные со стабилизацией экосистемы водохранилища (1961-1967 гг.), последствиями химизации сельского хозяйства (1968-1971 гг.), сукцессии (1972-1978 г.г.) и дальнейшего роста влияния техногенного фактора. Вторая главная компонента связана с уменьшением популяционной плотности каляноид, которое не было в полной мере учтено первой компонентой.



Показа-тели	ГК 1	ГК 2
Каляноиды	-0.282	**0.901**
Ротатории	**-0.773**	-0.446
Азот NH_4	**0.688**	-0.244
:Железо	**0.768**	0.100
Собствен-ные значения	**1.74**	1.08
Накоплен-ный % вариации	43.5	70.6

Рис. 7.16. Динамика изменения первой (сплошная линия) и второй (пунктир) главных компонент, обобщающих 4 показателя качества воды в Куйбышевском водохранилище; красная линия – линейный тренд 1-й главной компоненты

К разделу 7.4:

```
M8 <- read.table("time_6.txt",header=TRUE) ; M4 <- M8[, c(1,5,3,4)]
month = rep(5:10,nrow(M4)%/%6) ; M4.s <- transform(scale(M4),month = month)
#  1. Получение сезонной траектории фазового портрета
M4.ss <- M4.s[order(M4.s$month),] ; attach(M4.ss)
plot(M4.ss[,c(5,1)], type="n", ylim=c(-1,1.5), xlab="Номер месяца", ylab="Содержание отн.")
for (i in 1:4)  {
    lines(month, predict(loess(M4.ss[,i]~month), data.frame(month=month)),lwd=2, col = i)}
legend ("topright", c("Каляноиды",  "Ротатории", "Азот NH4", "Железо"),col = c(1:4),lwd=2)
# Функции конвертации временного ряда в таблицу лагов
lag.ts.matrix <- function(ts,order) {
       n <- length(ts);  x <- ts[(order+1):n]
```

```
       for (lag in 1:order) { x <- cbind(x,ts[(order+1-lag):(n-lag)])}
       colnames(x) <- c("lag0",paste("lag",1:order,sep=""))
       return(as.data.frame(x)) }
# Функция анализа главных компонент
# Не будем использовать rda() или prcomp(), а определим свою функцию
pca <- function(x, scale=T, center=T, proj=T) {
       nobs <- nrow(x);  nvars <- ncol(x); x <- scale(x, scale=scale, center=center)
       s <- svd(x/sqrt(nobs-1)) ; pca.lst <- list(Evals=s$d^2, Loadings=t(t(s$v)*s$d))
       if(proj) pca.lst$Projections <- t(t(s$u)*s$d)*sqrt(nobs-1)
       names(pca.lst$Evals) <- paste("Лямбда ",1:nvars,sep="")
       dimnames(pca.lst$Loadings) <- list(colnames(x),paste("Нагрузка ",1:nvars,sep=""))
       if(proj) dimnames(pca.lst$Projections) <-
                     list(rownames(x),paste("Компонента ",1:nvars,sep=""))
       for(j in 1:nvars)
         if(sum(pca.lst$l[,j]) < 0    { pca.lst$Loadings[,j] <- -pca.lst$Loadings[,j];
          if(proj) pca.lst$Projections[,j] <- -pca.lst$Projections[,j] }  ;  pca.lst }
# Формирование фазового портрета каляноид
L5.CAL <- lag.ts.matrix(M4[,1],5) ; pca.l <- pca(L5.CAL); pca.l$Loadings
pca.l$Evals; cumsum(pca.l$Evals[1:6]/sum(pca.l$Evals))
plot(pca.l$Projections[,c(1,2)], type="n" )
lines(pca.l$Projections[,c(1,2)]) ;
text(pca.l$Projections[,c(1,2)],labels= 6:108, cex=0.7, font=4)
#  2. Анализ многомерного тренда  методом ГК
W.decomp <- ts(M4,frequency = 6,start = c(1961,1)) ; M4.t <- M4
for (i in 1:4) {M4.t[,i] <-  decompose(W.decomp[,i])$trend}
pca.t <- pca(M4.t[4:105,]) ; pca.t$Evals; cumsum(pca.t$Evals[1:4]/sum(pca.t$Evals))
plot(M8[4:105,2], pca.t$Projections[,1], type="l", lwd=2)
abline(lm(pca.t$Projections[,1]~M8[4:105,2]),col="red", lwd=2)
lines(M8[4:105,2], pca.t$Projections[,2], lty=2)
```

7.5. Анализ пространственных структур

Под *пространственным анализом* (Spatial analysis) понимается набор методов исследований, в которых случайная переменная Z связана с некоторым изучаемым показателем, изменяющимся в пространстве и характеризующим динамику экосистемы или среды, а остальные независимые переменные определяют пространственное расположение объекта или точки наблюдения (X-Y координаты и высота H). Изучение пространственных структур играет очень важную роль при обработке результатов мониторинга (если не сказать категоричнее – большинство экологических исследований являются частным случаем пространственного анализа). В конечном итоге пространственно-временные технологии моделирования экосистем в различных масштабах рассматриваются как путь интегрального анализа и обобщения всего массива данных о состоянии природы и общества для обеспечения адекватных действий человечества по поддержанию среды и социума.

Теоретические основания анализа временных и пространственных рядов, такие как оценка автокорреляции или моделирование тренда, в целом имеют много общего, однако переход от вариации в одномерном градиенте времени t к изменчивости относительно двух-трех ортогональных пространственных осей существенно осложняет используемые математические процедуры. Легко убедиться в том, что пространственный анализ требует отдельного развернутого изложения, поэтому в настоящем пособии мы ограничимся лишь иллюстрацией некоторых узловых подходов.

При изучении пространственных структур принято выделять два направления: *геостатистический анализ* и *анализ пространственного размещения точек*. Отличие между ними заключается в способе реализации выборочного процесса. В первом случае изучаемое пространственно-распределенное явление рассматривается как случайная функция $Z(x)$, т.е. бесконечное множество случайных величин, представляющих некий непрерывный феномен в каждой точке пространства (Савельев и др., 2012). Исследователь

для анализа пространственного поведения этого феномена (например, содержания тяжелых металлов в верхнем слое почвы) сам планирует точки, в которых будут проводиться наблюдения, а полученные выборочные значения пространственной переменной $z(x)$ рассматриваются как одни из возможных реализаций функции $Z(x)$.

Во втором случае анализируются свойства пространственного рисунка и поведение явлений, которые проявляются дискретно в виде событий, но генерируются независимо от наблюдателя неким природным "механизмом" (Baddeley, 2010). Этот механизм, как и в первом случае, может быть представлен случайной функцией $Z(x)$, непрерывно распределенной в пространстве. О свойствах этого процесса мы также можем судить по выборочным данным, полученным в точках, однако их местоположение устанавливается не по воле исследователя, а случайно назначается в процессе генерации событий самим процессом (например, число яиц в гнезде можно посчитать только там, где есть гнездо). При этом предполагается, что у исследователя имеется полная карта координат всех точек, чтобы проанализировать их размещение.

В качестве примера (П7) рассмотрим изменчивость пространственной структуры популяций наземных моллюсков на участке 12×30 м с сеткой 8×20 пробных площадок. Всего было обнаружено 613 особей улиток *M. carthusiana* (из них, 281 ювенильных) и 352 экземпляра *B. Cylindrica* (в том числе, 262 ювенильных). Наш анализ с использованием пакетов статистической среды R во многом повторяет расчеты, сделанные в работе (П7 – Винарский и др., 2012) на основе программы PAST 2.16.

Одной из основных задач изучения пространственной структуры является визуальный анализ и графическое представление пространственной неоднородности распределения популяций с помощью карт локальных плотностей. Выполнить это можно с использованием различных алгоритмов сглаживания, например, на основе непараметрической ядерной регрессии (см. разделы 4.7 и 7.1). При этом скользящее "окно" сканирует всю область исследований и суммирует вклады точек в окрестности, связанной с текущим положением курсора. Гладкость полученной поверхности интерполяции зависит от ширины окна (bandwidth), которая может быть предварительно оценена по эмпирическим формулам или найдена автоматически в ходе кросс-проверки. Карта распределения популяционной плотности улиток *M. carthusiana* (рис. 7.17а) была построена с использованием изотропической гауссовой ядерной функции со стандартным отклонением $\sigma = 1.5$ м, которое играет здесь роль ширины окна (хотя в точном смысле таковой не является).

Важной дополнительной возможностью является построение карт пространственно-обусловленного относительного риска. Например, на рис. 7.17б представлена карта распределения доминирующих возрастных групп *M. Carthusiana*, позволяющая легко выделить фрагменты территории, где $p = n_{juv} / n_{ad} > 0.5$, т.е. число ювенильных особей n_{uv} превышает число взрослых животных n_{ad} (или наоборот). В этом примере речь, конечно, идет не о "риске" в привычном понимании, но если использовать, скажем, соотношение численностей больных и здоровых деревьев в лесу, то такая карта приобретет форму реального руководства к проведению природоохранных мероприятий.

Краеугольным камнем анализа пространственного распределения явлений является диагностика характера процесса, генерирующего события. В качестве нулевой гипотезы обычно принимается предположение о полностью случайном размещении точек в изучаемой области (CSR – Complete Spatial Randomness), моделируемом, например, однородным точечным процессом Пуассона (Bivand et al., 2008). Для проверки согласия анализируемых данных и гипотетической модели необходимо выбрать подходящую описательную статистику и сравнить ее эмпирическое значение с теоретической величиной. Типичными критериями здесь могут быть, например, распределение расстояний между ближайшими точками или среднее число "соседей" в круге фиксированного радиуса с центром в произвольной точке.

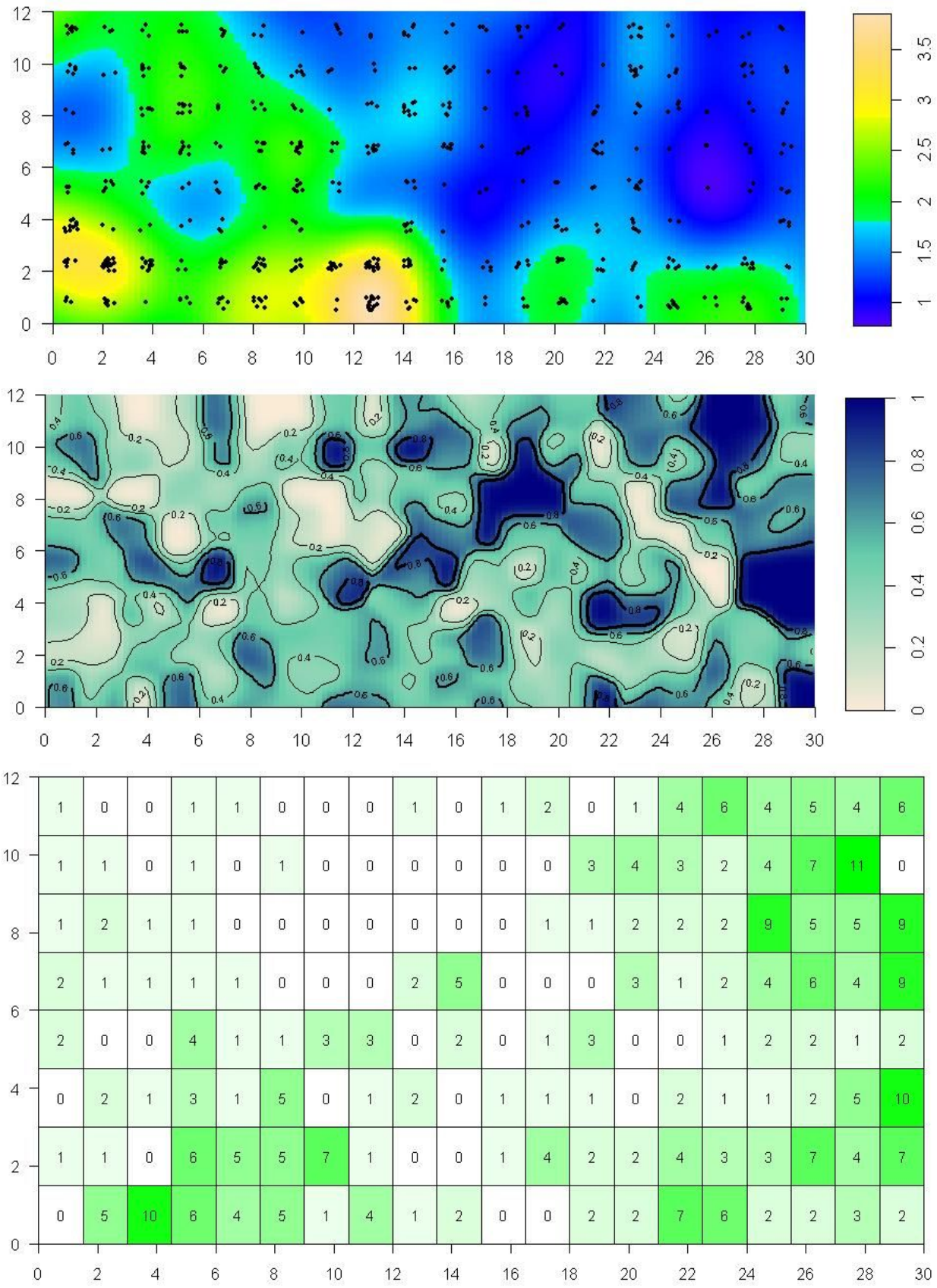
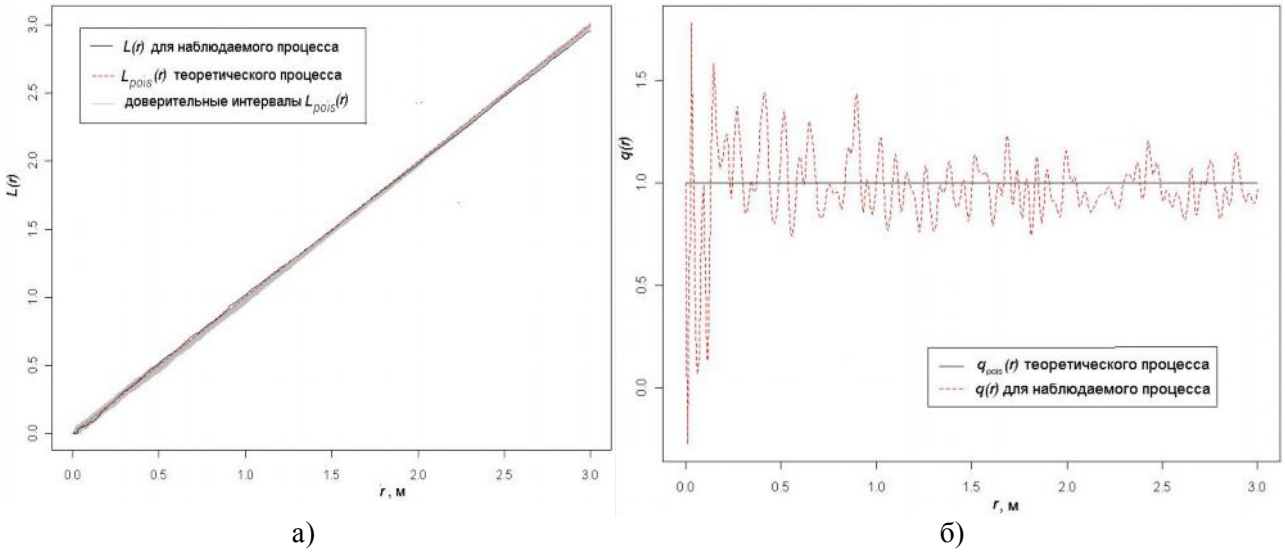


Рис. 7.17. Карты для анализа пространственной структуры популяций моллюсков: локальная ядерная плотность численности *M. carthusiana* (а); распределение соотношения между ювенильными и взрослыми экземплярами (б); размещение числа особей *B. Cylindrica* по сетке квадратов, соответствующей расположению пробных площадок (в)

Пусть r –радиус окружности, изменяющийся с некоторым шагом от 0 до $0.25w_{\min}$ ($w_{\min}$ – минимальный из двух размеров области исследования A). Тогда можно определить целую коллекцию различных функций распределения взаимных расстояний между точками (Baddeley, 2010) и выполнить построение соответствующих графиков:

- $G(r)$ – кумулятивной функции, основанной на распределении расстояний r между ближайшими соседями и пропорциональной числу таких дистанций, превышающих r; для однородного процесса Пуассона с интенсивностью λ: $G_{pois}(r) = 1 - e^{-\lambda\pi r^2}$;
- $K(r)$ –функции Рипли (или меры момента второго порядка), имеющей фактически тот же смысл, что и $G(r)$, но не зависящей от средней плотности точек; для пуассоновского процесса: $K_{pois}(r) = \pi r^2$;
- $L(r)$ – трансформированной функции Рипли $L(r) = [K(r)/\pi]^{0.5}$; $L_{pois}(r) = r$;
- $g(r)$ – функции парной корреляции точек $g(r) = K'(r)/\pi r$;
- $F(r)$ – функции "пустых пространств", основанной на распределении расстояний r от случайных координат окна, не содержащих точек, до ближайшей точки и другие.

Поскольку на рис. 7.17 можно усмотреть нестационарность размещения популяций улиток, для рассматриваемого примера нами использовалась версия процедур Kinhom(…) и Linhom(…) пакета spatstat, учитывающая неоднородность композиций точек. Однако следует оговориться, что расчет функций распределения расстояния между точками (рис. 7.18а, б) носит здесь исключительно демонстрационный характер, т.к. координаты расположения отдельных особей внутри каждой пробной площадки не были определены в эксперименте и моделировались нами как равномерный случайный процесс размещения. Поэтому не удивительно, что функции $G(r)$ и $K(r)$, вычисленные в пределах двух смежных квадратов, практически совпадали с аналогичными кривыми пуассоновского процесса. Предметные выводы, которые могут быть сделаны с помощью функции Рипли $K(r)$ на реальных примерах анализа пространственной структуры хвойно-широколиственных сообществ деревьев, изложены, например, в диссертации Н.А. Чижиковой (2008).



а) б)

Рис. 7.18. Графики функций для анализа распределения расстояний между особями моллюсков *M. carthusiana*: а)$L(r)$-функция и б) функция парной корреляции $q(r)$

Оценка статистической значимости $G(r)$- и $K(r)$-функций выполняется с помощью имитационных процедур Монте-Карло, генерирующих размещение объектов в соответствии с выбранной нулевой гипотезой. В этом случае используется метод огибающих оболочек (*simulation envelope test*), который является вариантом теста Барнарда (Barnard, 1993). Для этого рассчитывается B модельных кривых, соответствующих нулевому распределению, и для каждого значения r оцениваются

крайние границы, за которые эти кривые не могут выйти. Нулевая гипотеза отвергается, если кривая функций, вычисленная для наблюдений (например, картированного размещения улиток по территории), не будет полностью лежать между нижней и верхней огибающими. Поскольку тестируемая статистика является уже не скалярной величиной, а функцией, то процедура проверки гипотез сталкивается здесь с двумя проблемами (Грабарник, 2011): (а) как вычислить степень отклонения эмпирической статистики от теоретической кривой, соответствующей нулевой гипотезе и (б) следует ли использовать метод Бонферрони для множественного тестирования.

Перекрестная (cross-type) версия *K*-функции Рипли анализируемого многовидового процесса размещения точек определяется как $\lambda_v K_{uv}(r)$ и равна ожидаемому числу точек вида *v* (сверх случайного их числа), расположенных на расстоянии *r* от каждой точки вида *u*. Если размещение точек вида *u* независимо от размещения точек вида *v*, то значение $K_{uv}(r)$ равнялось бы r^2. Отклонение эмпирической кривой K_{uv} от теоретической кривой r^2 предполагает эффект взаимного влияния особей сравниваемых видов при их пространственном размещении – рис. 7.19.

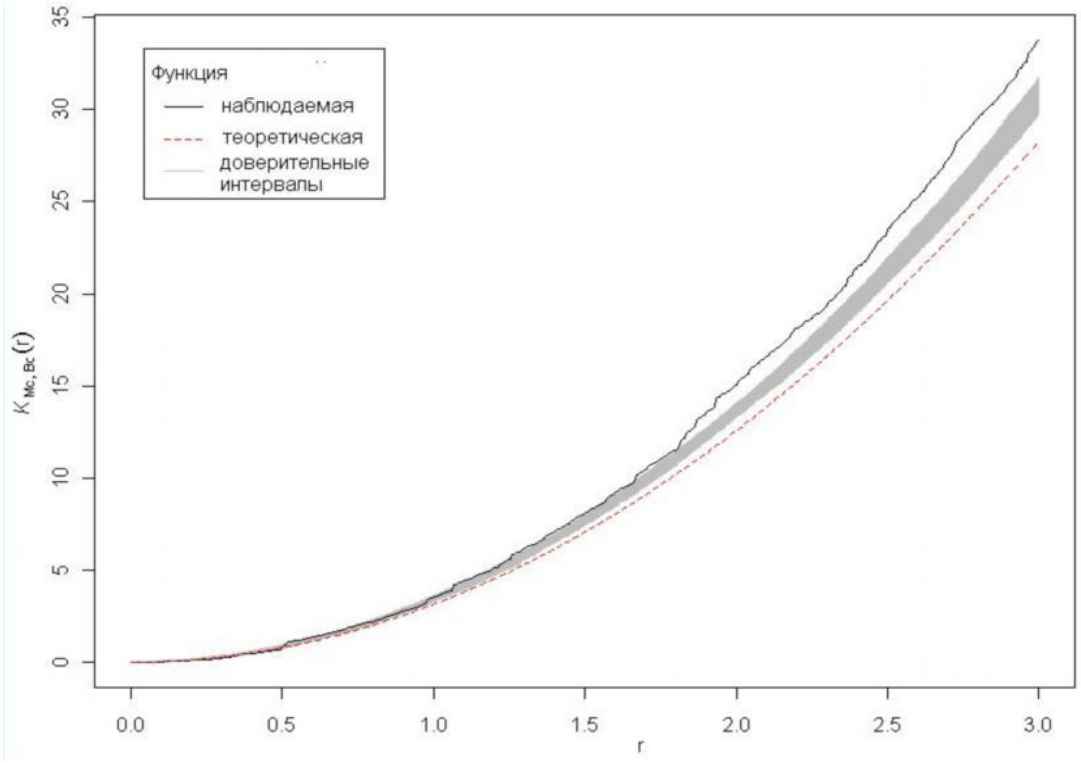


Рис. 7.19. График перекрестной функции Рипли для совместного размещения улиток видов *M. Carthusiana* и *B. Cylindrica*

Как отмечалось выше, случайность пространственного распределения точек оценивается сопоставлением наблюдаемых значений используемой функции и значений теоретической функции, генерируемой имитационными процедурами Монте-Карло. Наиболее распространена имитация однородного пуассоновского процесса, однако для оценки мощности критерия или обоснования пространственных моделей часто проводится имитация размещения, соответствующего выбранной альтернативе. Например, класс размещений, в которых точки образуют группы, моделируется процессами ‘Томас’ или Неймана-Скотта. Для этого вначале в изучаемую область *A* помещается n_R точек-"родителей", координаты которых случайны и независимы. На втором этапе формируется "дочерний процесс", т.е. в окрестностях каждого родителя согласно распределению Пуассона генерируется случайное количество потомков.

Другим способом оценить характер распределения популяционной плотности является агрегирование числа событий обнаружения на сетке с выбранным

разрешением и визуальная оценка размещения числа особей по квадратам (рис. 7.17в). При этом решается задача классификации наблюдаемого ансамбля реализаций, т.е. отнесение конкретной совокупности объектов к одному из возможных типов размещения: а) регулярное (детерминированное равномерное или стохастическое равномерное); б) групповое или кластеризированное (детерминированное неравномерное); в) неоднородное случайное и, наконец, г) полностью случайное размещение CSR. Для проверки гипотезы H_0 о соответствии конкретной конфигурации теоретической схеме размещения могут быть использованы как общеизвестные критерии χ^2 Пирсона или Колмогорова-Смирнова, так и специальные статистики, предложенные разными авторами: индексы Мориситы, Ллойда, Тейлора-Ивао, критерии Холгейта, Хопкинса, Бесага-Гливза, Кокса-Льюиса, Эберхарда, Диггля и другие (Миркин, Розенберг, 1978; Грабарник, Комаров, 1980).

Классическая форма теста на CSR основана на подсчете числа точек $n_{\text{obs}}(k)$, попавших в каждый k-й квадрат, k = 1, 2, …, m, сетки, наложенной на область исследования A, где $w = W_A / m$ – площадь каждого квадрата. При справедливости гипотезы о случайности размещения эмпирические частоты статистически значимо не отличаются от теоретических частот распределения Пуассона $Np_k = N(\lambda^k e^{-\lambda})/k!$, где N – общая численность всех точек. Несмещенной оценкой интенсивности процесса λ является среднее число точек на единицу площади: $\lambda = \bar{n}_{pois}(k)/w$. Проверка гипотезы о равенстве эмпирических и теоретических вероятностей H_0: $p = p_{\text{pois}}$ осуществляется, например, с использованием критерия согласия χ^2. В нашем примере для популяции *M. Carthusiana* было получено $\chi^2 = 352$ ($p < 0.0001$) и для *B. Cylindrica* $\chi^2 = 428$ ($p < 0.0001$).

Если в тесте χ^2 нулевая гипотеза не отклоняется, то проверяется второе предположение, заключающееся в оценке однородности (стационарности) пуассоновского процесса: число точек в любых двух непересекающихся частях области A являются случайными и независимыми переменными. Пространственная версия теста Колмогорова-Смирнова позволяет проверить, равномерно ли распределено число точек вдоль осей x_1 или x_2. Сделанные расчеты D-статистики позволяют сделать вывод, что размещение популяций изучаемых моллюсков не является ни случайным, ни стационарным:

в горизонтальном направлении		в вертикальном направлении;	
M. Carthusiana	*B. Cylindrica*	*M. Carthusiana*	*B. Cylindrica*
$D = 0.267$ ($p < 0.0001$)	$D = 0.125$ ($p < 0.0001$)	$D = 0.155$ ($p < 0.0001$)	$D = 0.16$ ($p < 0.0001$)

Представленные критерии обладают необходимой мощностью только при достаточной репрезентативности выборок, что требует большого числа пробных площадок. Поэтому часто используется индекс Морисита (Morisita, 1959, здесь и далее до конца раздела цит. по Винарский и др., 2012 – П7): $IM = \dfrac{m\sum_{i=1}^{m} n_i(n_i - 1)}{n(n-1)}$,

где m – количество пробных площадок; n_i – количество особей в пределах i-того квадрата; n – общее количество особей на всех площадках. Как утверждают Винарский с соавторами (2012), на величину индекса Морисита не оказывают ощутимого влияния ни размеры квадратов, ни представительность выборки, что является далеко не очевидным.

При случайном типе размещения индекс Морисита IM = 1 и имеет верхний и нижний 95% доверительные интервалы, рассчитываемые по формулам (Hairstone et al., 1971): $M_l = \dfrac{\chi^2_{0,025} - m + n}{n-1}$; $M_u = \dfrac{\chi^2_{0,975} - m + n}{n-1}$.

Другая возможность построить доверительные интервалы индекса Мориситы при справедливости случайного размещения состоит в использовании имитационных моделей. Будем многократно (B = 1000) воспроизводить однородный пуассоновский процесс и генерировать случайные распределения точек по квадратам исследуемой области с одной и той же интенсивностью. Если на каждой итерации вычислять значение IM, то легко найти оценки M_l и M_u по полученному распределению, например, методом процентилей.

Вывод о типе пространственного распределения делается следующим образом:

если $IM \le M_l$ – распределение равномерное;
если $M_l < IM < M_u$ – распределение случайное;
если $IM \ge M_u$ – распределение групповое (агрегированное или нестационарное).

В нашем примере для всех подпопуляций моллюсков может быть принята гипотеза о групповом характере их обитания – табл. 7.2. Оценки доверительных интервалов, рассчитанные бутстрепом и по формулам Хаирстона с соавторами совпали с точностью до трех значащих цифр и отдельно не выделялись.

Таблица 7.2. Оценки пространственного размещения популяций улиток с использованием индекса Мориситы *IM* и показателя внутривидового агрегирования Айвза *J* ; M_l и M_u - доверительные интервалы индекса *IM* при справедливости нулевой гипотезы, J_{jack}, J_l и J_u - оценка "складного ножа" для *J* и ее доверительные интервалы

Вид, возраст	M_l	*IM*	M_u	*J*	J_{jack}	J_l	J_u
B. cylindrica, ad	0,629	***2,277***	1,414	1,271	***1,270***	0,532	2,008
B. cylindrica, juv.	0,873	***1,769***	1,141	0,771	***0,777***	0,389	1,166
B. cylindrica (все)	0,906	***1,766***	1,105	0,769	***0,769***	0,507	1,031
M. carthusiana, ad	0,900	***1,433***	1,111	0,434	***0,441***	0,129	0,753
M. carthusiana, juv.	0,882	***1,293***	1,131	0,294	0,303	-0,033	0,639
M. carthusiana (все)	0,946	***1,316***	1,060	0,318	***0,321***	0,136	0,506

А.Айвз (Ives, 1988, 1991) предложил показатель внутривидовой агрегации (intraspecific aggregation), который является модификацией индекса дисперсии: $J = (S^2 / D - 1) / D$, где D и S – средняя плотность и оценка ее дисперсии для анализируемого вида. Значение *J* обычно интерпретируется как доля конспецифичных особей к ожидаемому их числу при случайном размещении. Следовательно оценки показателя, достоверно не отличающиеся от нуля, свидетельствуют о пуассоновском характере процесса распределения особей в пространстве против альтернативы о их агрегированности.

Для оценки статистической значимости гипотезы H_0: $J = 0$ нам в очередной раз предоставляется убедительная возможность применить методы ресамплинга. Используя алгоритм "складного ножа" (jackknife), убираем из наших данных численность особей первого квадрата и на основе оставшихся рассчитываем величину J_{-1}. Повторяем эту процедуру для каждого квадрата из используемого размещения и в итоге получаем псевдовыборку, содержащую *m* значений J_{-i}. На ее основе мы можем получить оценки складного ножа для всех основных статистических параметров:

- jackknife-оценку искомого показателя: $J_{jack} = m \cdot J - \left[\frac{m-1}{m}\right]\sum_{i=1}^{m} J_{-i}$, где J – значение показателя внутривидовой агрегации, рассчитанное по полной выборке;
- оценку ошибки показателя *J*: $S_J^* = \left[\frac{m-1}{m}\sum_{i=1}^{m}(J_{-i} - \overline{J}_{-i})^2\right]^{0.5}$; $\overline{J}_{-i} = \left(\sum_{i=1}^{m} J_{-i}\right)/m$;
- нижнюю и верхнюю границы 95% доверительного интервала *J*:
  - $J_l = J_{jack} - t_{0.05}\, S_J^*$ и $J_u = J_{jack} + t_{0.05}\, S_J^*$ соответственно,

где $t_{0,05}$ – табличное значение критерия Стьюдента для уровня значимости α = 0,05 и числа степеней свободы $df = m - 1$.

В нашем примере для ювенильных особей вида *M. carthusiana* показатель внутривидовой агрегации Айвза, оценивая размещение улиток как случайное, вступил в противоречие с индексом Мориситa (табл. 7.2). Последнего также поддержали результаты теста статистики χ^2, которые показали, что число особей, приходящееся на единицу площади, нельзя считать постоянным: $\chi^2 = 241$, $p = 0.0003$.

Другой показатель межвидовой агрегации (interspecific aggregation) популяций со средними плотностями D_1 и D_2, также предложенный Айвзом, оценивает меру увеличения (уменьшения) числа гетероспецифичных особей по отношению к ожидаемому их числу при случайном распределении: $C_{12} = Cov_{12}/(D_1 \cdot D_2)$, где Cov_{12} – ковариация численностей в квадратах между анализируемой парой видов 1 и 2.

Показатель межвидовой агрегации C будет иметь положительный знак, если два вида имеют позитивную ассоциативность, и отрицательный – в случае конкуренции за пространственный ресурс. В нашем примере для видов *B.cylindrica* и *M. carthusiana* (объединено для всех возрастных групп) этот индекс имеет значение C = -0,13, что свидетельствует о том, что при увеличении численности одного вида плотность другого вида имеет тенденцию к уменьшению.

Для того, чтобы оценить, насколько статистически значимо показатель Айвза отличается от нуля, которому соответствует отсутствие межвидового взаимодействия, опять воспользуемся методами ресамплинга. Доверительные интервалы величины C при справедливости нулевой гипотезы можно найти по тому же алгоритму, что мы использовали для индекса Мориситы, т.е. с применением моделей, имитирующих случайное совместное распределение анализируемых видов. Диапазон критических значений $C_{крит}$, ограничивающих область H_0, составил от -0.053 до 0.0593, т.е. величина показателя C для наших эмпирических данных является статистически значимой.

К разделу 7.5:

```
MOL <- read.xls("Mol.xls", sheet = 1, rowNames=FALSE) ; attach (MOL)
library(spatstat)  #  Расчеты проводим на базе пакета spatstat
#  Две функции, осуществляющие случайное размещение особей внутри пробной площадки
q.point <- function(xq, yq, nq) {
       xp <- matrix(rep(0, nq*2), ncol=2) ; xp[,1] <- 1.5*(xq - 1) + runif(nq,0,1.5)
       xp[,2] <- 1.5*(yq - 1) + runif(nq,0,1.5)  ; return (xp) }
df.point <- function(df, cx, cy, cm) {
       df.nz <- subset(df, df[,cm]>0, select = c(cx,cy,cm))
       x <- q.point(df.nz[1,1],df.nz[1,2], df.nz[1,3])
       for (i in 2:nrow(df.nz)) {x <- rbind(x, q.point(df.nz[i,1],df.nz[i,2], df.nz[i,3]))}
       colnames(x) <- c("X","Y") ; return( transform(as.data.frame(x),
                     spec=substr(colnames(df)[cm],1,2), age=substr(colnames(df)[cm],4,6))) }
# Создадим объекты ppp (point pattern)
p.Mc <- df.point(MOL,cm=5,cx=1,cy=2); p.Mc <-  rbind(p.Mc, df.point(MOL,cm=6,cx=1,cy=2))
ppp.Mc <- ppp(x=p.Mc$X, y=p.Mc$Y, marks=data.frame(spec=p.Mc$spec, age=p.Mc$age ),
                                                  window=owin(c(0, 30), c(0, 12)))
# вычислим размер скользящего окна bandwidth по правилу  Silverman
sigma<-(sd(ppp.Mc$x)+sd(ppp.Mc$y))/2 ; iqr<-(IQR(ppp.Mc$x)+IQR(ppp.Mc$y))/2
bandwidth <- 0.9*min(sigma, iqr)*ppp.Mc$n^(-1/5)
Mc.intensity <- density.ppp(ppp.Mc, sigma=bandwidth)
plot(Mc.intensity) ; points(ppp.Mc, pch=19, cex=0.6)  #  Получаем рис. 7.16а
axis(side=1, at = seq(0, 30, 2), las = 1, pos = 0)   # ось снизу
axis(side=2, at = seq(0, 12, 2), las = 1, pos = 0)   # слева
ppp.Temp <- ppp.Mc ; marks(ppp.Temp) <- ppp.Mc$marks$age   # выделим коды возрастов
p <- relrisk(ppp.Temp, 0.5) # вычислим вероятность появления особей juv для рис. 7.16б
plot(p,col=colorRampPalette(c("antiquewhite", "aquamarine3","navyblue"))(100))
contour(p,nlevels=5, lwd=seq(from=0.1, to=3, length.out=5),add=T)  # добавим изолинии
p.Bc <- df.point(MOL,cm=3,cx=1,cy=2); p.Bc <-  rbind(p.Mc, df.point(MOL,cm=4,cx=1,cy=2))
ppp.Bc <- ppp(x=p.Bc$X, y=p.Bc$Y, marks=data.frame(spec=p.Bc$spec, age=p.Bc$age ),
                  window=owin(c(0, 30), c(0, 12)))
# построим сетку ячеек
ppp_qu<-quadrats(ppp.Bc, nx = 20, ny = 8) ; xgrid <-  ppp_qu$xgrid; ygrid <-  ppp_qu$ygrid
quadcount <- quadratcount(ppp.Bc, tess=ppp_qu) # посчитаем число точек в ячейках
image(xgrid, ygrid,      t(quadcount[order(1:8, decreasing = T), ]),
                      col = colorRampPalette(c("white", "green"))(15),axes=F)
plot(quadcount, add=T, cex=0.9)   #  Получаем рис. 7.16в
#добавим точки  plot(ppp.Bc, which.marks="age", chars=c(19, 24), cex=0.5, add=T)
# Получение функций K и L расстояния между точками
```

```
Mc.Kfv <- envelope(unmark(ppp.Mc), fun=Kinhom) ; plot(Mc.Kfv); plot(pcf(Mc.Kfv))
Mc.Lfv <- envelope(unmark(ppp.Mc), fun=Linhom) ; plot(Mc.Lfv)
# Получение K-функции  для парного размещения видов
df.all <-  rbind(p.Mc[,-4], p.Bc[,-4])  ; ppp.all <- ppp(x=df.all$X, y=df.all$Y,
                    marks=data.frame(spec=df.all$spec), window=owin(c(0, 30), c(0, 12)))
plot(Kcross.inhom(ppp.all))
all.Cfv <- envelope(ppp.all, fun=Kcross.inhom, method="b") ; plot(all.Cfv)
#  Хи-квадрат тест CSR с использованием численности точек в квадратах
quadrat_test_result <- quadrat.test(ppp.Bc, nx = 20, ny = 8,alternative="clustered")
 quadrat_test_result$observed             # Наблюдаемое число экземпляров в квадрате
round(quadrat_test_result$expected , 2)   # Ожидаемое число экземпляров
quadrat.test(ppp.Mc[ppp.Mc$marks$age=="juv"],nx = 20, ny = 8) #   Для группы особей
kstest(ppp.Bc, "x") ; kstest(ppp.Bc, "y")   #  Тест Колмогорова-Смирнова
# Сравнение индекса Мориситы с его распределением при справедливости CSR
Morisita <- function (N, n=length(N)) {
       NS <- sum(N) ; IM <- n*sum(N*(N-1))/NS/(NS-1) # Индекс Мориситы
       DI <- (qchisq(c(0.025,0.975),(n-1))+NS-n)/(NS-1) # Его доверительные интервалы
       return (list(IM=IM, DI=DI)) }
# Нахождение ДИ бутстрепом. Задаются вектор численностей, размеры окна и сетки
IM.randtest <- function( N, lx, ly,  nx, ny, Nperm) {
       PermArray <- as.numeric(rep(NA, Nperm)) ; lambda <- sum(N)/lx/ly
    for (i in 1:Nperm) {pp <- rpoispp(lambda,win=owin(c(0,lx),c(0,ly))) # размещение CSR
      quadcount <- quadratcount(pp, nx=nx, ny=ny)
      PermArray[i] <- Morisita(as.data.frame(quadcount)$Freq)$IM }
    return (RandRes(Morisita(N)$IM, PermArray, Nperm)) }
source("print_rezult.r")  # Загрузка функций вывода результатов
IM.randtest((MOL$Mc_ad+MOL$Mc_juv), 30, 12, 20, 8, 1000)
J.index <- function(x){(var(x)/mean(x)-1)/mean(x)}  # Функция расчета индекса J Айвза
library(bootstrap)   #  Оценка доверительных интервалов J  методом складного ножа
N <- MOL$Mc_juv ; results <- jackknife(N,J.index)
J.index(N); results$jack.bias ; results$jack.se
tc <- qt(0.975, length(N)-1)  # Оценка доверительных интервалов по Jackknife-оценкам
Jl <- J.index(N) - results$jack.bias - tc* results$jack.se ;
Ju <- J.index(N) - results$jack.bias + tc* results$jack.se
# Функции расчета индекса C  Айвза и оценки его рандомизированных доверительных интервалов
C.index <- function(N1, N2){cov(N1, N2)/mean(N1)/mean(N2)}  # Функция расчета C-индекса
C.randtest <- function( dfa, nx, ny, Nperm) {
       PermArray <- as.numeric(rep(NA, Nperm)) ; types <- names(dfa) ; nal <- colSums(dfa)
       for (i in 1:Nperm) { ppp <- rmpoint(nal, 1, types=types ) # размещение CSR
                quadcount.N1 <- quadratcount(ppp[ppp$marks==types[1]], nx=nx, ny=ny)
                quadcount.N2 <- quadratcount(ppp[ppp$marks==types[2]], nx=nx, ny=ny)
                PermArray[i] <- C.index(as.data.frame(quadcount.N1)$Freq,
                                        as.data.frame(quadcount.N2)$Freq) }
       return (RandRes(C.index(dfa[,1], dfa[,2]), PermArray, Nperm)) }
dfa <- data.frame(Bc=MOL$Bc_ad+MOL$Bc_juv, Mc = MOL$Mc_ad+MOL$Mc_juv)
C.randtest(dfa, 20, 8, 1000) #  Расчет индекса C Айвза и оценка его значимости
save(MOL, file="MOL.RData")
```

7.6. Автоковариация и пространственно обусловленная зависимость отклика

Важным является замечание, что представленные ранее индекс Мориситы, показатели Айвза или χ^2 -статистика опираются на среднее соотношение дисперсии и оценки плотности популяций и фактически не учитывают изменчивость особей по пробным площадкам как функцию расстояния между ними. Пространственная структура, заключенная в выборочной матрице отклика **Z**, может формироваться под влиянием двух основных причин: (а) под воздействием внешних (экологических) факторов, которые в свою очередь пространственно структурированы, и/или (б) непосредственно как результат пространственной дифференциации процессов внутри самого сообщества. В первом случае говорят о пространственно обусловленной зависимости, во втором – о пространственной автокорреляции.

Модель пространственно обусловленной зависимости (induced spatial dependence) предполагает, что значение $z(\mathbf{x}_0)$ переменной отклика в точке с координатами $\mathbf{x}_0$ равно:

$$z(\mathbf{x}_0) = \mu_z + f(\mathbf{X}_0) + \varepsilon(\mathbf{x}_0) \qquad \text{(Borcard et al., 2011)}$$

где μ_z – общая средняя переменной z, $\mathbf{X}_0$ – совокупность независимых переменных, и ε – некоррелированные остатки, которые случайно варьируют с изменением пространственных координат. Поле переменных $\mathbf{X}$ образует детерминированную структуру (градиент), непосредственно определяющую модель пространственной изменчивости случайной величины $z(\mathbf{x})$.

В реальности пространственно обусловленная зависимость приводит к феномену локальных сгущений (агрегаций) и разрежений плотности распределения экологических объектов, т.е. образованию кластеров – см. рис. 7.20 для примера [П7].

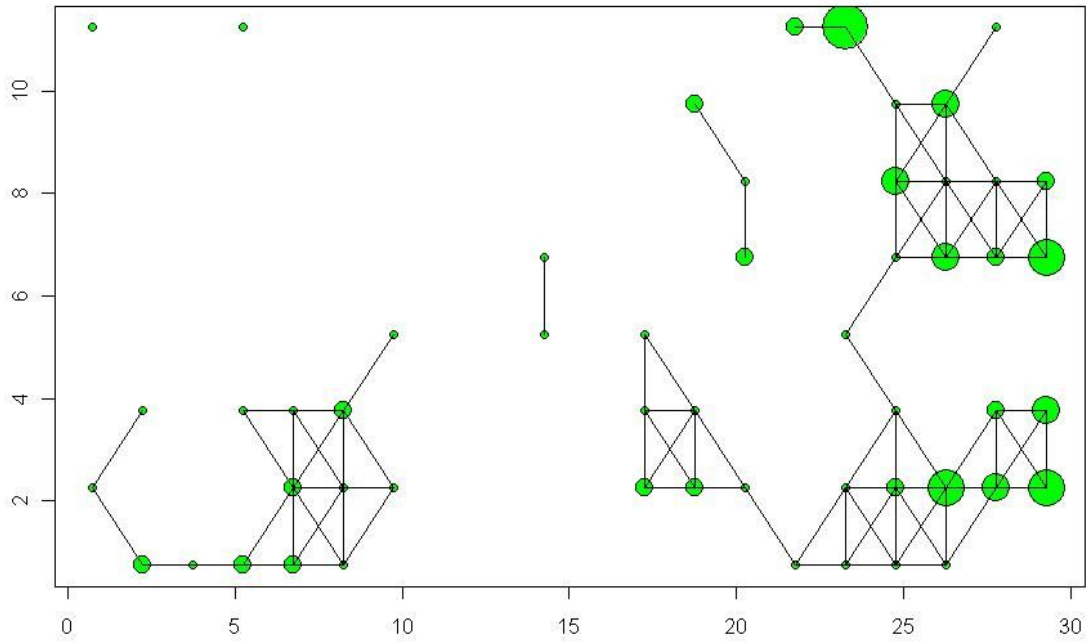


Рис. 7.20. Сгущения взрослых особей моллюсков *B.cylindrica*; линиями соединены соседние участки, где эти улитки были представлены, а радиус кругов пропорционален численности обнаруженных экземпляров

Если образование кластеров связано с нарушениями независимости выборочного процесса (иногда обнаруженные сгущения "провоцируют" исследователя выполнять повторный отбор проб именно на этих участках), то их наличие приводит к смещенным оценкам параметров генеральной совокупности – среднего, дисперсии, вариограммы. Поскольку использование большинства геостатистических методов основаны на предположениях о стационарности, эргодичности и мультинормальности распределения пространственной переменной $z(\mathbf{x})$, то такая утрата независимости наблюдений нарушает эти предположения. Функция клеточной декластеризации для среды R, осуществляющая расчет весов, приписываемых каждому элементу выборки в условиях локальных сгущений, приводится в пособии (Савельев и др., 2012).

Прогноз значений исследуемой переменной $z(x)$ в каждой точке координат двухмерной сетки может быть осуществлен путем построения моделей пространственного тренда с использованием, например, традиционных функций нелинейной регрессии. Выполним аппроксимацию распределения улиток по пробным площадкам на основе полного полинома 3-й степени. Выбор информативного комплекса предикторов может быть сделан на основе любой процедурой селекции из числа описанных в разделе 4.6. Однако мы хотим обратить внимание читателя на эффективный алгоритм прямого поиска (Blanchet et al., 2008), реализованный в функции forward.sel(…) пакета packfor и использующий критерий "двойного останова". Он обеспечивает максимум приведенного

коэффициента детерминации $\overline{R}^2$ при статистической значимости всех предикторов модели, оцениваемой по рандомизационному тесту.

Трехмерный график пространственного тренда оценок численностей моллюсков *B.cylindrica*, нормированных по формуле Хеллингера (см. раздел 5.3), представлен на рис. 7.21, а сама полученная модель регрессии при ($\overline{R}^2 = 0.237$ имеет вид:

$$\hat{n}_{Bc} = 0.73 - 0.037x + 0.0013x^2 + 0.00197xy - 0.045y.$$

Аналогичная модель для распределения математического ожидания числа особей *M. Carthusiana* по пробным площадкам может быть записана как

$$\hat{n}_{Mc} = 0.914 - 0.0102x - 0.00011xy \quad (\overline{R}^2 = 0.127).$$

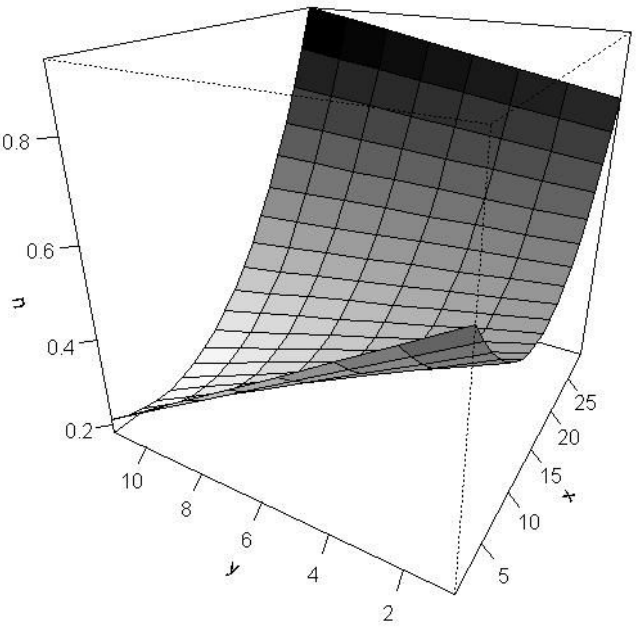


Рис. 7.21. Полиномиальная поверхность пространственного тренда математического ожидания нормированной численности моллюсков *B.cylindrica*

Модель пространственной автокорреляции переменной z в точке x_0:

$$z(\mathbf{x}_0) = \mu_z + \Sigma f[z(\mathbf{x}_h) - \mu_z] + \varepsilon(\mathbf{x}_0)$$

определяет зависимость отклика $z(\mathbf{x}_0)$ от значений z в точках, находящихся в окружности радиусом h и центром с координатами $\mathbf{x}_0$. Чем больше расстояние h, тем меньший вклад вносится сопряженными точками в оценивание пространственной переменной $z(\mathbf{x})$. При наличии автокорреляции данных значение отклика в произвольной точке может быть предсказано (по крайней мере, частично) по его выборочным наблюдениям в h-окрестности. На этой основе развивается геостатистический подход, известный как *кригинг* (Bivand et al., 2008), который «строит скорее статистическую модель реальности, чем модель интерполяционной функции» (Савельев и др., 2012, с. 17).

Пространственная автокорреляция отражает тот факт, что экологические объекты, находящиеся в относительной пространственной близости, более связаны между собой, чем случайно отобранные пары. Данное явление известно как закон Тоблера, сформулированный в полушутливой форме: «всё влияет на всё, но то, что ближе, влияет сильнее». Автокорреляция, которая является функцией расстояния между изучаемыми локациями, оценивается в геостатистике путем визуального анализа графиков различных структурных функций – коррелограмм, вариограмм и периодограмм –, а также с использованием строгих статистических тестов.

Распространенным способом выделения доли пространственной ковариации в общей вариации исходных данных является использование I-статистики Морана, которая подобно коэффициенту корреляции Пирсона варьирует в интервале от -1 до +1:

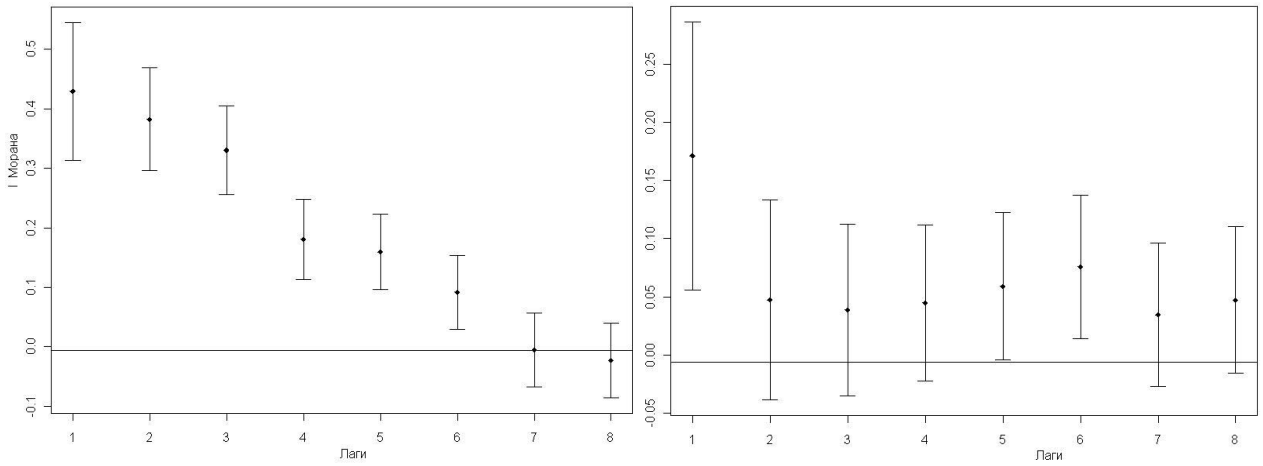
$$I = \left(\frac{m}{\sum_{i=1}^{m}\sum_{j=1}^{m} w_{ij}} \right) \cdot \left(\frac{\sum_{i=1}^{m}\sum_{j=1}^{m} w_{ij} \cdot (z_i - \bar{z}) \cdot (z_j - \bar{z})}{\sum_{i=1}^{m} (z_i - \bar{z})^2} \right),$$

где m – число точек или пространственных единиц (в нашем случае, пробных площадок); z_i – значение изучаемой переменной (численность улиток в i-м квадрате); $\bar{z}$ – среднее

значение признака для всей популяции; w_{ij} – "вес", который отражает степень пространственной близости между точками i и j (расстояние между каждой парой пробных площадок). При отсутствии пространственной автокорреляции математическое ожидание коэффициента Морана будет $E(I) = -1/(m-1)$. Стандартная ошибка выборочных значений и доверительные интервалы I могут быть найдены по формулам аппроксимации или методами Монте-Карло с использованием бутстрепа.

Коррелограмма представляет собой график значений пространственной автокорреляции в зависимости от лага h. Типичной является коррелограмма, в которой значения I положительны для небольших расстояний, уменьшаются по мере увеличения лага до отрицательных значений и стабилизируются в точке h_{crit}, правее которой пространственную структуру можно считать статистически независимой.

На рис. 7.22а, б приведены коррелограммы численности двух видов наземных моллюсков (Винарский и др., 2012). Для улиток *B.cylindrica* коэффициент Морана плавно уменьшается при увеличении лага, оставаясь статистически значимым фактически на всей ширине области исследования (6×1.5 = 9 м), что свидетельствует о наличии выраженного монотонного пространственного тренда. В частности, если выполнить расчет коррелограммы для остатков приведенной выше полиномиальной модели (т.е. для выборки с элиминированным трендом), то коэффициенты Морана для всего набора лагов оказываются статистически незначимыми.



а) б)

Рис. 7.22. Графики автокорреляционных функций на основе коэффициента I Морана и его доверительных интервалов для численности моллюсков *B.cylindrica* (а) и *M. Carthusiana* (б)

Для вида *M. Carthusiana*, напротив, подавляющее большинство оценок автокорреляции значимо не отклоняется от ожидаемой величины при справедливости H_0, что свидетельствует о формировании случайного паттерна при распределении особей в пределах изученного региона. Здесь также следует упомянуть о необходимости вносить поправку Бонферрони для критического уровня значимости, учитывающего множественность испытаний. В частности, для примера на рис. 7.22б доверительные интервалы коэффициента Морана при лагах 5 и 6 не включают значение 0, однако после выполненной коррекции Бонферрони их статистическая значимость оказывается равной 0.312 и 0.0613 соответственно.

Другим основным инструментом анализа пространственной связи является вариограмма. При условии справедливости гипотезы стационарности второго порядка вариация $\gamma(\mathbf{h})$ и ковариация $C(\mathbf{h})$ являются двумя равноценными мерами для характеристики взаимосвязи между случайными величинами $Z(\mathbf{x})$ и $Z(\mathbf{x} + \mathbf{h})$, разделенными в пространстве вектором $\mathbf{h}$: $\gamma(\mathbf{h}) = \mathrm{E}\{[Z(\mathbf{x} + \mathbf{h}) - Z(\mathbf{x})]^2\} = C(0) - C(\mathbf{h})$.

Эмпирическая вариограмма $\gamma^*(h)$ является выборочной оценкой функции $\gamma(h)$ и строится на основе данных наблюдений $\{z(\mathbf{x}_1), z(\mathbf{x}_2),\ldots, z(\mathbf{x}_n)\}$. Для заданной

фиксированной последовательности векторов **h** специальным образом выделяются подмножества пар точек $\{\mathbf{x}_i, \mathbf{x}_j\}$, находящихся на расстоянии, близком к $|\mathbf{h}|$, на основе которых вычисляется набор значений вариограмм (точнее полу- или семивариограмм):

$$\gamma^*(h) - \frac{1}{2N_h}\sum_{i=1}^{N_h}\left[z(x_i) - z(x_i + h)\right]^2 .$$

В общем случае график "правильной" вариограммы ведет себя следующим образом:

- начинается в нулевой точке, т.к. для $|\mathbf{h}| = 0$ $\quad z(\mathbf{x} + \mathbf{h}) - z(\mathbf{x}) = 0$;
- возрастает с ростом $|\mathbf{h}|$;
- стабилизируется на определенном уровне или неограниченно растет (см. пример на рис. 7.23).

В вариограммном анализе принято выделять ряд характерных геометрических феноменов эмпирической кривой. Величину C_0, соответствующую значению $\gamma^*(0)$ в нулевой точке, называют "эффектом самородка" (nagget) – см. рис. 7.23а-б. Если $C_0 > 0$, то это является признаком наличия микроструктур в масштабе меньше лага, либо связан с ограниченностью выборки при слишком большом расстоянии между точками. Прямая линия вариограммы с C_0 = const соответствует полному отсутствию корреляционной зависимости между точками, как бы близки они не были.

Чтобы проверить, нельзя ли объяснить изменение полувариограммы с увеличением лага h чисто случайными причинами, можно построить большой набор вариограмм с использованием тех же самых данных, но после многократного случайного перемешивания измерений относительно пространственных координат. Если эмпирическая вариограмма находится, например, в 95 % интервале изменчивости рандомизированных псевдовариограмм, то вполне вероятно предположение о полной пространственной хаотичности базового процесса. Впрочем, заключение о полной некоррелированности обычно весьма редко для реальных естественных процессов. Пример совмещенного графика эмпирической и рандомизированных вариограмм для распределения численности улиток по пробным площадкам представлен на рис. 7.23а и здесь гипотеза отсутствия пространственной корреляции кажется нам маловероятной.

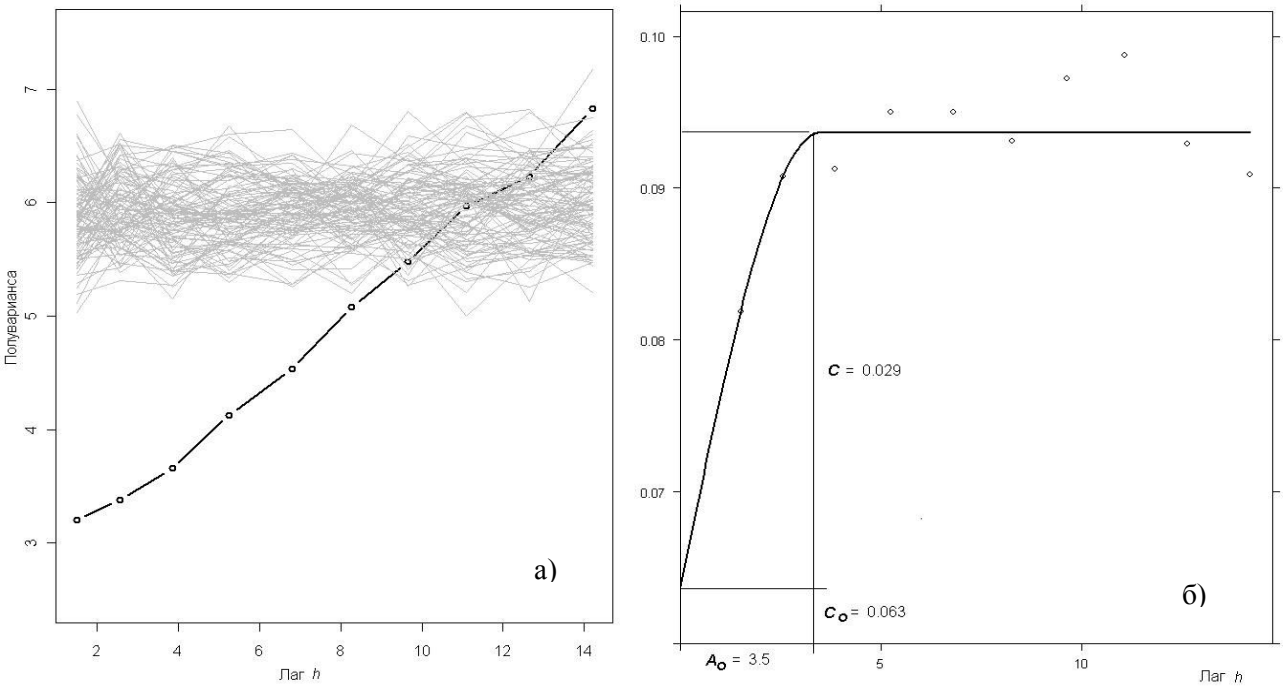


Рис. 7.23. Вариограммы абсолютной (а) и "детрендированной" (б) численности моллюсков *B.cylindrica* ; серым цветом показаны графики вариансы при справедливости нулевой гипотезы, буквенные обозначения – по тексту

Если график вариограммы растет линейно и неограниченно, то это связано обычно с наличием выраженного пространственного тренда при ограниченной области исследований. Вариограмма, построенная нами по тем же выборочным данным, но после элиминации тренда, имеет вид характерной параболы с насыщением (рис. 7.23б). Величина $C + C_0$ называется "порогом" (sill) и соответствует максимально возможному уровню изменчивости анализируемой величины. Наконец, величина A_0, равная значению лага, при котором вариограмма достигает своего порога, называется "радиусом пятна" (range), за пределами которого значения $Z(\mathbf{x})$ и $Z(\mathbf{x} + \mathbf{h})$ становятся некоррелированными.

Вариограммный анализ обычно преследует две основных цели:

- моделирование эмпирической вариограммы с использованием подходящей базисной функции, к числу которых относятся сферическая, экспоненциальная, гауссова или степенная функции, и с учётом эффекта "самородков" C_0 и величины "пятен" A_0;
- оценку анизотропности, т.е. анализ, как меняется характер вариограммы, если выполнять отсчет лага разделяющего вектора по различным географическим направлениям.

Процедуры подбора моделей, аппроксимирующих эмпирическую вариограмму, в целом используют общие принципы статистического моделирования: необходимо найти положительно определенную функцию, имеющую единственное устойчивое решение и доставляющую минимум ошибки по отношению к наблюдаемым данным. В частности, на рис. 7.23б нами была использована в качестве базовой широко употребляемая сферическая модель. В дальнейшем модели вариограмм подставляются в уравнения кригинга для получения искомых оценок изучаемой пространственной переменной (Савельев и др., 2012).

Эффективным способом визуальной оценки анизотропии изучаемого явления является построение поверхности вариограммы, представленной для рассматриваемого примера на рис. 7.24.

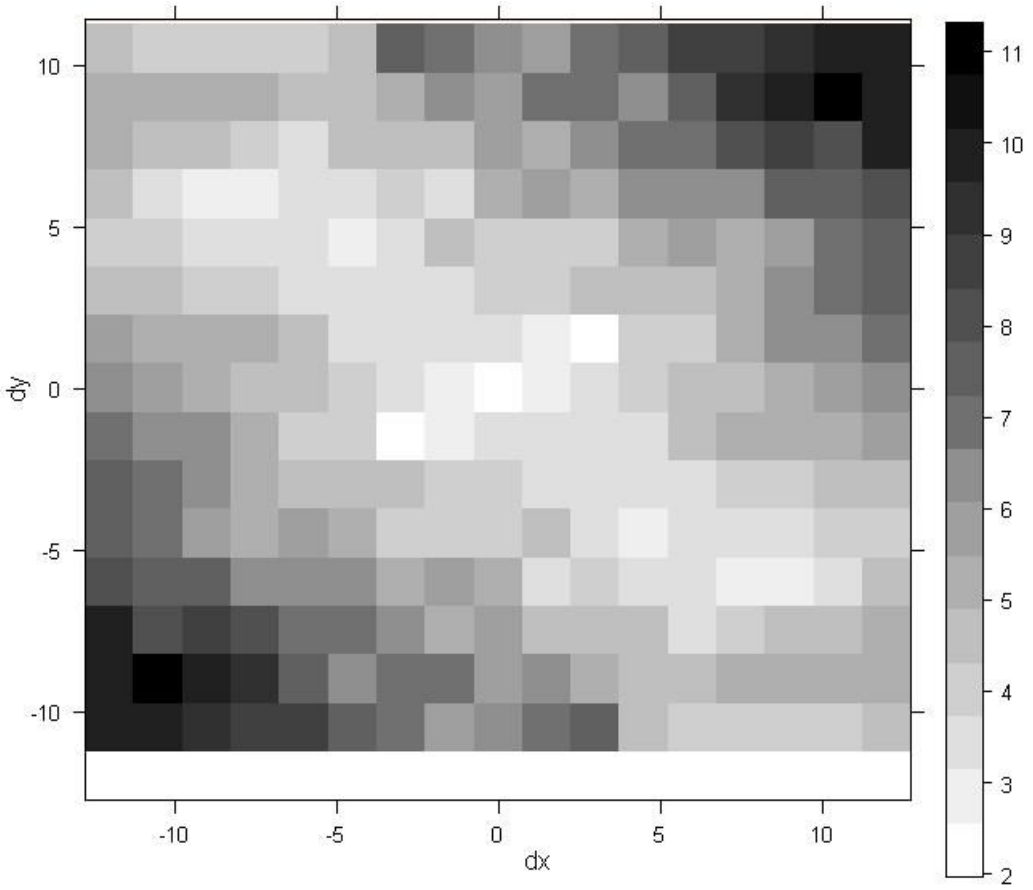


Рис. 7.24. Поверхность вариограммы численности моллюсков *B.cylindrica* ;

На построенной поверхности (рис. 7.24) хорошо просматривается неоднородность (анизотропия) пространственной связи по диагоналям квадрата, обусловленная выявленным ранее трендом, который представлен моделью на рис. 7.21. В направлениях 135^o и -45^o по азимуту вариограмма нарастает гораздо медленнее (а, следовательно, число моллюсков изменяется менее резко), чем в других направлениях. Это может быть объяснено, например, градиентом влажности или изменчивостью растительности. Дальнейшее исследование пространственной структуры заключается в подробном анализе графиков поведения набора экспериментальных вариограмм, построенных по заданным фиксированным направлениям связи (например, в направлении максимума анизотропии и перпендикулярном к нему).

Автокорреляция и пространственный тренд представляют самостоятельный интерес, однако становятся негативными факторами при исследовании и статистическом анализе влияния переменных окружающей среды на интенсивность биологических процессов. Они приводят к тому, что наблюдения случайных величин уже не являются независимыми, и каждый новое измерение не привносит с собой полной степени свободы. Например, при построении модели зависимости биопродуктивности растительного покрова от влажности случайность и независимость распределения ошибок может быть нарушено автокорреляцией, число степеней свободы окажется завышенным и нулевая гипотеза станет отклоняться чаще, чем это было бы при отсутствии пространственной гетерогенности.

Аналогичный эффект может иметь место и при изучении межвидовых взаимодействий, поэтому для оценки взаимной согласованности (coherence) или взаимного исключения (turnover) видов необходимо принимать во внимание условие стационарности второго порядка. При справедливости этой гипотезы вводится предположение, что математическое ожидание и дисперсия наблюдаемых случайных величин в совокупности не зависит от пространственных координат, а их автоковариация зависит только от лага. Поэтому для того, чтобы оценить наличие автокорреляции между многовидовыми структурами с использованием статистики Мантеля Z_M, которая была описана нами ранее в разделе 6.3, необходимо использовать матрицы остатков после элиминации пространственного тренда.

Пусть $\mathbf{d}_Z$ – матрица различий между каждой парой пробных площадок в многомерном пространстве видов, вычисленная с использованием любой из подходящих мер (например, нормированного расстояния Евклида или индекса Брея-Кёртиса). Тогда статистика Мантеля $Z_M = \mathbf{d}_Z\mathbf{d}_X$ оценивает корреляцию между экологическим сходством сообществ и географическими расстояниями ними $\mathbf{d}_X$. Если в матрице $\mathbf{d}_X$ выделить только пары соседних участков, а остальным элементам присвоить значение 0, то получим матрицу географических расстояний $\mathbf{d}_X^1$ с лагом 1. Тогда оценкой автокорреляции может быть статистика Мантеля для начального (первого) лага $Z_M^1 = \mathbf{d}_Z\mathbf{d}_X^1$. Процесс вычислений повторяется для всей последовательности лагов h = 2, 3, …, т.е. каждый раз строятся новые модельные матрицы $\mathbf{d}_X^h$ и рассчитываются значения статистики Z_M^h (Legendre, Legendre, 1998). Коррелограмма Мантеля представляет собой график изменения нормированных величин r_M^h в зависимости от классов h. Статистическая значимость r_M^h проверяется с использованием рандомизационного теста или путем аппроксимации статистики Мантеля нормальным распределением.

Если $\mathbf{d}_Z$ имеет смысл матрицы различий, то положительные значения статистики Мантеля соответствуют отрицательной автокорреляции между численностями видов для данного значения лага, как это имеет место в рассматриваемом примере при h = 1 на рис. 7.25 (т.е. два смежных квадрата на единичном расстоянии друг от друга будут иметь различающуюся видовую структуру). Другое статистически значимое отрицательное значение r_M^h при h = 7 соответствует положительной зависимости между обилиями между улиток *B.cylindrica* и *M. Carthusiana* на расстоянии около 12 м. Отметим также, что статистическая значимость статистики Мантеля для всех лагов оценивалась с

использованием 999 случайных перестановок, а коррекция наблюдаемых *p*-значений на множественность испытаний методом Бонферрони выполнялась последовательно таким образом, чтобы можно было бы обнаружить наличие автокорреляции прежде всего в первых классах расстояний.

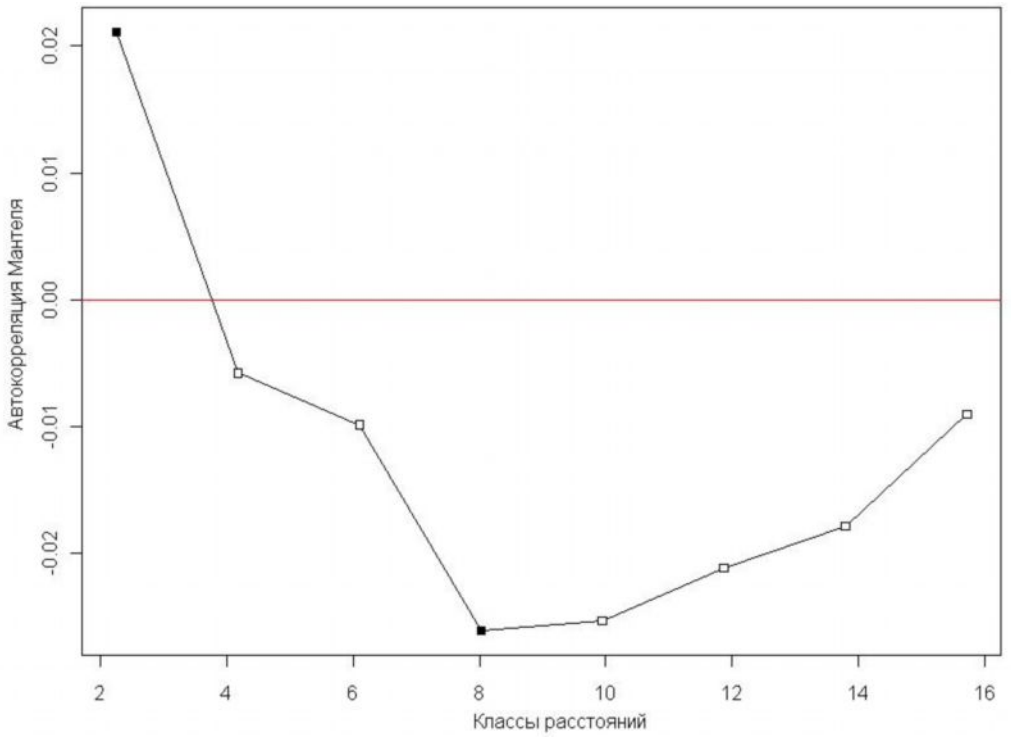


Рис. 7.25. Коррелограмма Мантеля на примере популяций улиток;
черным цветом закрашены статистически значимые значения статистики Мантеля по результатам 1000 итераций рандомизационного теста с учетом коррекции Бонферрони

К разделу 7.6:

```
load (file="MOL.RData");  library(spdep) ; library(vegan) ; library(gstat)
MOL.xy <- data.frame(x=1.5*(MOL[,1] - 1) + 0.75,  y=1.5*(MOL[,2] - 1) + 0.75)
MOL.sp <- data.frame(Bc = MOL[,3]+MOL[,4], Mc = MOL[,5]+MOL[,6])
#  Вывод графа связей между смежными пробными площадками, где обнаружены особи
MOL.Bc <- subset(cbind(MOL.xy, MOL[,3:4]), Bc_ad > 0) ; D.mat = as.matrix(dist(MOL.Bc[,1:2]))
plot(MOL.Bc[,1:2], type="p", pch=21, bg="green", cex=5*(MOL.Bc[,3])/max((MOL.Bc[,3])))
n = nrow(MOL.Bc); thresh = sqrt(2*1.5^2)   # Соединяем всех соседей в радиусе 2.12 м
for(j in 1:(n-1)) {  for(jj in (j+1):n)  {    if((D.mat[j,jj]<= thresh)&(D.mat[j,jj]>0))
             lines(c(MOL.Bc[j,1], MOL.Bc[jj,1]), c(MOL.Bc[j,2], MOL.Bc[jj,2])) } }
#  Построение моделей пространственного тренда
MOL.h <- decostand (MOL.sp, "hellinger") # Преобразование по Хеллингеру
MOL.poly <- poly(as.matrix(MOL.xy), degree=3, raw=TRUE)
colnames(MOL.poly) <- c("X","X2","X3","Y","XY","X2Y","Y2","XY2","Y3")
# Получение полной модели и последующая селекция информативных переменных
MOL.trend <- lm(as.matrix(MOL.h) ~ ., data=as.data.frame(MOL.poly)); summary(MOL.trend)
install.packages("packfor", repos="http://R-Forge.R-project.org")
forward.sel(MOL.h$Bc, MOL.poly, adjR2thresh=0.2313)
MOL.Bc.trend <- lm(MOL.h$Bc ~ X + X2 + XY + Y, data=as.data.frame(MOL.poly))
forward.sel(MOL.h$Mc, MOL.poly, adjR2thresh=0.1291)
MOL.Mc.trend <-lm(MOL.h$Mc ~ X + XY , data=as.data.frame(MOL.poly))
summary(MOL.Bc.trend)  ;  summary(MOL.Mc.trend)
#   Получение детрендированных остатков
MOL.h.det <- as.data.frame(cbind(resid(MOL.Bc.trend),resid(MOL.Mc.trend)))
# Функция создания палитры фаций для трехмерного графика
hgt.pal <- function(z) {ny=ncol(z); nx=nrow(z)
       hgt <- 0.25 * (z[-nx,-ny] + z[-1,-ny] + z[-nx,-1] + z[-1,-1])
```

```
      hgt <- (hgt - min(hgt))/ (max(hgt) - min(hgt)) ; return (hgt) }
z <- matrix(MOL.Bc.trend$fit, ncol=8, nrow=20)
persp(unique(MOL.xy$x),unique(MOL.xy$y), z, theta = -60, phi = 25, xlab = "x",
            ylab = "y", zlab = "n", ticktype = "detailed", col = gray(1 - hgt.pal(z)))
#  Построение вариограммы для вида B.cylindrica по натуральным данным
MOL.v <- cbind(MOL.xy,MOL.sp) ; coordinates(MOL.v)=~x+y
Bc.v <- variogram(Bc ~ 1, MOL.v, cutoff=15, width=1.5)
plot(Bc.v$dist,Bc.v$gamma, ylim=c(2.5,7.5), type="b", lwd=2)
for (i in 1:100) {   #  Линии вариограммы  для рандомизированного набора данных
      Bc.vr <- variogram(sample(MOL.v$Bc,nrow(MOL.v)) ~ 1, MOL.v, cutoff=15, width=1.5)
      lines(Bc.vr$dist,Bc.vr$gamma, col="grey") }
Bc.v <- variogram(Bc ~ 1, MOL.v, cutoff=12, width=1.5,map=TRUE) ;
plot(Bc.v,col.regions=gray((16:0)/16)) #  Отрисовка поверхности  вариограммы
#  Построение вариограммы и ее модели по данным после элиминации тренда
MOL.dt.v <- cbind(MOL.xy, MOL.h.det) ; coordinates(MOL.dt.v)=~x+y
Bc.dt.v <- variogram(BcDt ~ 1, MOL.dt.v, cutoff=15, width=1.5)
 Mc.mv <- vgm(0.08, "Sph", 5, nug = 0.02)  # Параметры предполагаемой модели
plot(Bc.dt.v, model = fit.variogram(Bc.dt.v, model = Bc.mv),
                                                col=1, ylim=c(0.06,0.105), lwd=2)
#  Построение коррелограмм на основе коэффициента Морана
nb1 <- dnearneigh(as.matrix(MOL.xy), 0,1.5)
Bc.correlog <- sp.correlogram(nb1, MOL.sp$Bc, order=8, method="I",zero.policy=TRUE)
Bc.det.cor <- sp.correlogram(nb1, as.vector(MOL.h.det[,1]),
                                                order=8, method="I", zero.policy=TRUE)
print(Bc.correlog, p.adj.method="bonferroni") ; print(Bc.det.cor, p.adj.method="bonferroni")
Mc.correlog <- sp.correlogram(nb1, MOL.sp$Mc, order=8, method="I",zero.policy=TRUE)
plot(Bc.correlog) ; plot(Mc.correlog)
MOL.h.D1 <- dist(MOL.h.det); # Построение коррелограммы Мантеля
(MOL.man_cor <- mantel.correlog(MOL.h.D1, XY=MOL.xy, nperm=999)) ; plot(MOL.man_cor)
MOL.man_cor$n.class ; MOL.man_cor$break.pts
```

7.7. Байесовский подход и марковские цепи Монте-Карло

С именем Томаса Байеса (или в правильной фонетической транскрипции Бейза – Bayes, 1763) связывается широкий спектр различных методов оценки статистических параметров, схем принятия решений и алгоритмов распознавания образов. Иногда название “байесовский подход” намекает на использование формулы Байеса для условных вероятностей, но чаще является просто синонимом слова “вероятностный”, хотя и в определенном нетривиальном контексте. Тем не менее, в общих чертах отличие байесовской парадигмы от классической статистической теории можно определить достаточно четко (McCarthy, 2007, Айвазян, 2008; Link, Baker, 2009):

1. Классическая статистика оперирует *частотным* определением “вероятности”, под которой понимается предел отношения определенного результата эксперимента к общему числу экспериментов. *Байесовское* понимание вероятности – это степень уверенности в истинности суждения, а теорема Байеса задает правило, по которому эта степень уверенности изменяется при появлении новой информации.

2. Байесовский подход использует любую предварительную информацию относительно формулируемой гипотезы или параметров модели, которая обычно выражается в виде *априорной* функции вероятности предполагаемого результата. Затем начальные предположения "пересматриваются" с учетом эксперимента или выборочных данных, что находит свое отображение в виде уточненного *апостериорного* распределения плотности вероятности.

3. Байесовские методы стремятся ответить на логически состоятельный вопрос: "Какова вероятность того, что H_0 верна, если принять во внимание полученные данные?". Оцениваемые параметры трактуются уже не как некие предположительные константы, а как *случайные величины*, и результаты в виде графиков их вероятностных распределений в полной мере отвечают поставленной задаче и легко интерпретируются. Напомним, что

классическая проверка гипотез отвечает на весьма косвенный вопрос "Какова вероятность получить наши данные при условии, что H_0 верна?" (и это при том, что данные уже получены, а аналогичная вероятность при H_1 не всегда может быть определена).

Байесовское правило определяет соотношение между априорными вероятностями $p(A_i)$, определяющими уровень наших знаний до проведения эксперимента (или до обработки имеющихся данных), и апостериорными условными вероятностями $p(A_i|D)$, скорректированными по результатам эксперимента D:

$$p(A_i \mid D) = p(A_i)\,p(D \mid A_i) / \sum_i p(A_i)\,p(D \mid A_i)\,.$$

Здесь A_i , $i = 1, ..., k$ - набор гипотез (моделей, теорий, суждений) об изучаемом предмете. Эти гипотезы являются взаимоисключающими и образуют исчерпывающее множество возможных объяснений изучаемого феномена $\sum_i p(A_i) = 1$. Функция п*равдоподобия* $p(D|A_i)$ соответствует вероятностям того, насколько результаты эксперимента подтверждают правильность i-й гипотезы. Поскольку знаменатель приведенной формулы не зависит от параметров эксперимента, соотношение Байеса показывает, как меняются априорные представления о предмете в процессе накопления знаний:

$$\{\text{начальные знания}\} + \{\text{данные эксперимента}\} \underset{\text{модель}}{\Longrightarrow} \{\text{конечные знания}\}.$$

Оцениваемый параметр может иметь дискретное распределение, т.е. принимать несколько фиксированных значений с вероятностями, соответствующими каждой гипотезе. Рассмотрим пример трансформации после двух испытаний представлений эколога о классе качества воды в изучаемой реке: A_1 – река чистая, A_2 – река "грязная". Пусть априорные оценки вероятностей этих состояний, основанные, например, на общих представлениях эколога о реках региона равны $p(A_1)$=0.3 и $p(A_2)$=0.7. Были взяты гидробиологические пробы и сделан расчет индекса EPT, точность которого 80% (т.е. в 20% случаев положительного теста река будет ошибочно квалифицироваться как чистая). Тогда после первого испытания апостериорные оценки вероятностей будут пересчитаны так, что при положительном результате теста:

p(чистая| EPT+) = 0.3·0.8/(0.3·0.8 + 0.7·0.2) = 0.632 ; p(грязная| EPT+) = 1 - 0.632 = 0.368,

а если поденок, веснянок и ручейников в реке не оказалось, то

p(чистая| EPT-) = 0.3·0.2/(0.3·0.2 + 0.7·0.8) = 0.097 ; p(грязная| EPT-) = 1 - 0.097 = 0.903.

При повторном положительном тесте EPT апостериорные вероятности становятся априорными: p(чистая| EPT_2+) = 0.632·0.8/(0.632·0.8 + 0.368·0.2) = 0.873,

а при достаточном числе испытаний окончательный вывод будет уже мало зависеть от первоначальных предположений.

Однако чаще считается, что случайная величина параметра θ распределена непрерывно с некоторой плотностью $p(\theta)$, $p(\theta) \geq 0 \;\forall\; \theta$, $\int_\Theta p(\theta)d\theta = 1$. Если нет иных веских предположений, то априорную вероятность параметра можно считать распределенной равномерно. Имея выборку наблюдений и рассчитав функцию правдоподобия, можно найти условную плотность распределения параметра при данной выборке – см. рис. 7.26. Математическое ожидание случайной величины, имеющей такую условную плотность, называется байесовской оценкой параметра. Область **C**, такая, что $\int_C p(\theta)d\theta = 1-\alpha$, $1 \geq \alpha \geq 0$, называется (1 - α)-ой доверительной областью параметра, а ее граничные значения – интервалом высокой апостериорной плотности (Highest Posterior Density, HPD).

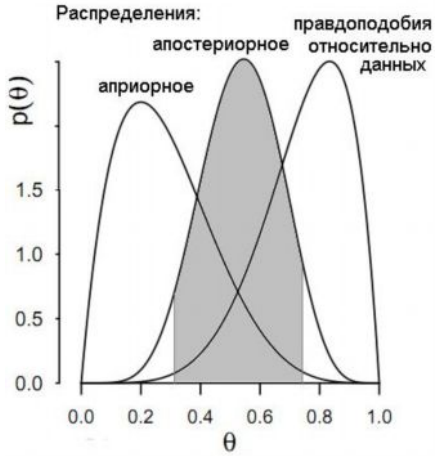


Рис. 7.26. Иллюстрация байесовских распределений; серым цветом выделена область высокой апостериорной плотности

Рассмотрим первый пример, связный с оценкой метеорологических явлений. По данным за 10-летний период, представленным норвежскими учеными (T. Reitan, http://folk.uio.no/trondr), из n = 3652 дней наблюдений в k = 596 случаях шел дождь. При этом в k_2 = 393 случаях он начинался на следующий день после солнечной погоды, но в k_1 = 202 случаях продолжался и на следующий день. Модель 1 предполагает, что вероятность дождя p в каждый день является равновероятным событием, зависящим только от общей частоты его появления в наблюдаемом ряду, и оценивает правдоподобие относительно наблюдаемых данных D как:

$$\Pr_1(X_1, X_2, ..., X_n) = \sum\nolimits_p \Pr(D \mid p) \Pr(p) = p^k (1 - p)^{n-k} = 0.0154,$$

где X_i – результат наблюдения в i-й день, в нашем примере равный 0 (нет дождя) или 1.

Модель 2 основывается на предположении, что появление дождя зависит от ближайшей предыстории и оценивает две вероятности: p_1 = Pr("дождь"|"дождь в предыдущий день") и p_2 = Pr("дождь"|"нет дождя в предыдущий день"). Таким образом, принимается, что условная вероятность события "дождь сегодня" зависит только от вероятности вчерашнего дождя и не зависит от событий более раннего периода. Это соответствует простой цепи Маркова с дискретным временем

$$\Pr_2(X_1, X_2, ..., X_n) = \Pr(X_1) \cdot \Pr(X_2 \mid X_1) \cdot \Pr(X_3 \mid X_2) \cdot ... \cdot \Pr(X_n \mid X_{n-1}).$$

Если также принять естественное предположение о стационарности процесса и считать, что вероятности остаются неизменными для любых пар i – (i - 1), то правдоподобие модели 2 относительно наблюдаемых данных D можно рассчитать как

$$\Pr_2(X_1, X_2, ..., X_n) = \sum\nolimits_{p_1, p_2} \Pr(D \mid p_1, p_2) \Pr(p_1) \Pr(p_2) = p_0 \, p_1^{k1} (1 - p_1)^{k-k1} \, p_2^{k2} (1 - p_2)^{n-k-k2},$$

где p_0 – вероятность дождя в любой из дней, $p_0 = p_2 / (1 - p_1 + p_2)$.

Сравнить степень правдоподобия обоих моделей по отношению к данным можно с использованием байесовского K-фактора, который является определенной альтернативой классическим средствам проверки гипотез. В нашем случае

$$K = \Pr_1(X_1, X_2, ..., X_n) / \Pr_2(X_1, X_2, ..., X_n) = 2.32 \cdot 10^{-29},$$

т.е. модель 2 гораздо лучше согласуется с анализируемыми данными, чем модель 1.

Распределение апостериорных плотностей вероятностей p_1 и p_2 для модели 2 представлено на рис. 7.27. С использованием полученных распределений можно при необходимости рассчитать математические ожидания и доверительные интервалы этих параметров, оценить корреляцию θ между вероятностями этих двух событий и др.

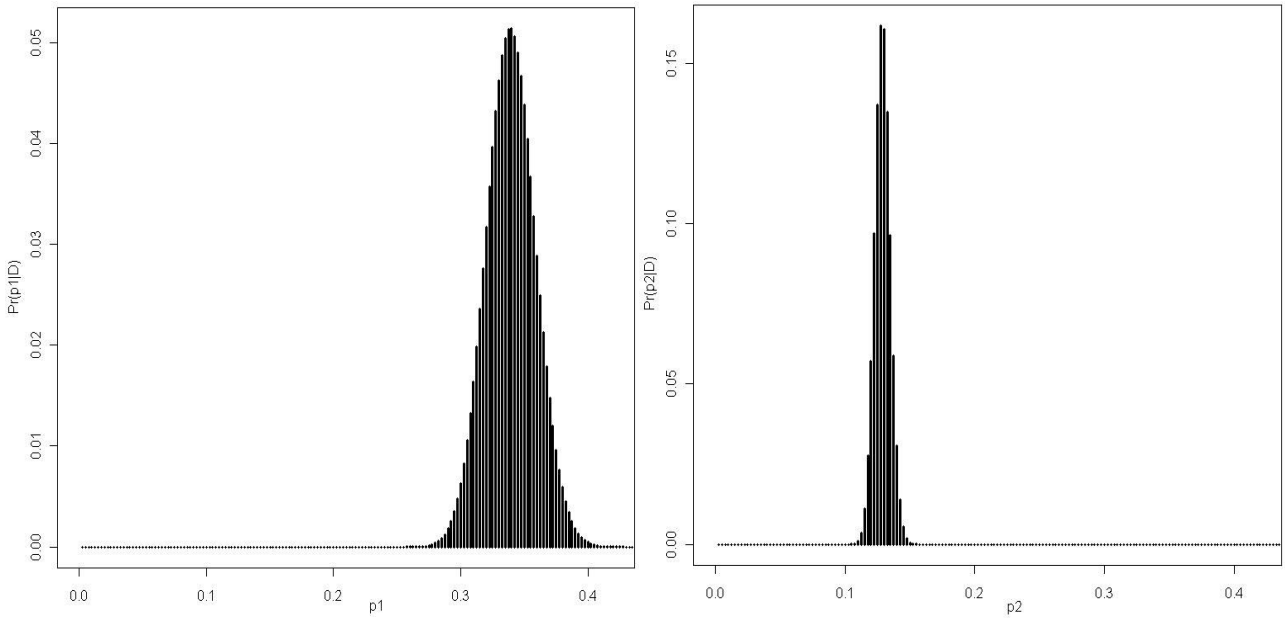


Рис. 7.27. Распределение плотности апостериорных вероятностей дождя при наличии (p_1) и отсутствии (p_2) такового в предшествующий день

Интересно также построить распределение разности плотности вероятности вероятностей "вторичного" и "первичного" дождя (p_1 - p_2) и оценить соотношение между этими параметрами. Для нашего примера среднее значение (p_1 - p_2) = 0.208, а вероятность справедливости гипотезы $\Pr(p_2 > p_1)$ крайне мала и равна $6.8 \cdot 10^{-32}$.

В более сложных случаях для получения байесовских оценок параметров и их апостериорных распределений необходимо выполнять многократное генерирование случайных величин с заданным распределением. Эффективными средствами генерации таких выборок являются итерационные методы Монте-Карло, использующие цепи Маркова (MCMC – Monte Carlo Markov chain), разработку которых иногда трактуют как наиболее существенный прорыв в статистике за последние несколько десятков лет.

Основой моделирования с помощью MCMC служит построение марковского процесса, для которого стационарное распределение переходов определяется функцией $P(\theta|X)$. Процесс моделирования довольно длительный: конструируется большое количество марковских цепей с заданными параметрами до тех пор, пока распределение текущих значений не приблизится к стационарному распределению переходов. Существует ряд алгоритмов конкретной реализации процесса моделирования – генерирование выборки по Гиббсу, алгоритм Метрополиса-Хастингса и др. (Andrieu et al., 2003; Seefeld, Linder, 2007; Бидюк и др., 2009; Hartig et at., 2011).

Метод Гиббса (Gibbs sampler) представляет собою способ формирования выборок ($\mathbf{x}_1$; …; $\mathbf{x}_n$) из заданных распределений $p(x)$ m-мерных переменных путем применения многоразовых выборок из одномерных условий:

- вначале инициализируется начальный вектор $\mathbf{x}_0 = (x_1^0,...,x_m^0)$;
- на первой итерации $i = 1$ генерируется последовательность случайных величин: x_1^1 – из распределения $p(x_1 \mid x_2^0, x_3^0,..., x_m^0)$, x_2^0 – из $p(x_2 \mid x_1^0, x_3^0,..., x_m^0)$, …, x_m^1 – из $p(x_m \mid x_1^0, x_2^0,..., x_{m-1}^0)$ и т.д.; в результате получаем $\mathbf{x}_1 := (x_1^1,...,x_m^1)$;
- процесс порождения случайных величин повторяется достаточно большое число раз $i = 2, \ldots, n$, $n > 10000$, чтобы цепь успела достигнуть сходимости к своему стационарному распределению.

Если n велико, то результирующий набор значений $\mathbf{x}_i := (x_1^i,...,x_m^i), i = \overline{1,n}$, в совокупности будет совпадать с общим распределением $p(x)$. Поскольку марковский процесс не может сразу стабилизироваться, то на практике некоторую часть случайных значений на начальных итерациях считают образцами для испытаний на "розжиг процесса" (brun-in), в окончательный набор не включают и используют лишь для оценки близости генерируемых значений к общему распределению.

Генератор Гиббса применим, когда есть возможность сформировать выборки из полного условного распределения. В отличие от него алгоритм Метрополиса-Гастингса (Metropolis-Hastings MCMC sampler) используется, например, в случаях, когда точечные реализации распределений можно получать, задавая вектор искомых параметров **Φ**. В итерационном процессе участвует два распределения, одно из которых π является "априорным" (proposal), т.е. задающим амплитуду "скачков" **Φ** в зависимости от необходимой точности расчетов. Другое распределение $L(\mathbf{\Phi})$ – это целевое апостериорное распределение параметров, которое мы восстанавливаем в ходе имитаций.

Алгоритм в общих чертах состоит из следующих шагов:

° принимается конкретный вид пропорционирующего распределения π вероятностей перехода в цепи Маркова и определяется вид процедуры, возвращающей апостериорное распределение $L(\mathbf{\Phi})$ с учетом априорных вероятностей и функции правдоподобия;

° инициируется начальный вектор параметров $\mathbf{\Phi}_0$ и рассчитываются исходные значения цепи Маркова $C_1 \Leftarrow L(\mathbf{\Phi}_0)$;

° с использованием распределения π текущий вектор $\mathbf{\Phi}$ модифицируется в $\mathbf{\Phi}^*$ и вычисляется отношение апостериорных плотностей $R = p(\mathbf{\Phi}|\text{ данные})/p(\mathbf{\Phi}^*|\text{ данные})$;

° в зависимости от значения R следующее звено цепи C_{i+1} составят либо величины $\mathbf{\Phi}^*$ (итерация принимается и $\mathbf{\Phi} \Leftarrow \mathbf{\Phi}^*$ – зеленые кружки на рис. 7.28), либо остаются значения цепи C_i, найденные на предыдущем шаге (модификация отклоняется – красные кружки);

° делается большое число (10000-100000) циклов наращивания цепи C и набор ее значений представляет собой точечную аппроксимацию распределения $L(\mathbf{\Phi})$.

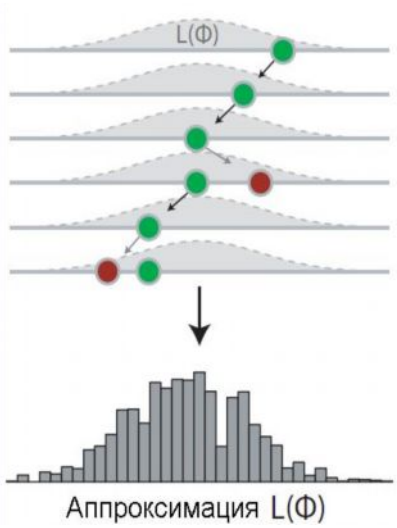


Рис. 7.28. Иллюстрация работы алгоритма Метрополиса-Гастингса, выполняющего точечную аппроксимацию неизвестной функции апостериорного распределения $L(\mathbf{\Phi})$ искомых параметров $\mathbf{\Phi}$

Рассмотрим второй пример, связанный с оценкой параметров простой регрессионной модели. Пусть необходимо проанализировать зависимость длины тела L ящерицы прыткой *Lacerta agilis* без учета хвоста от массы особи M (П5). Классическая аллометрия обычно базируется на предположении Хаксли (Huxley), считающего, что две произвольные фенотипические характеристики животных x и y связаны между собой показательным уравнением $y = \beta_1 x^{\beta 2}$. Поэтому простейший путь вычислений – это выполнить логарифмическую трансформацию выборочных данных и рассчитать обычным путем уравнение линейной регрессии: $\ln(\hat{L}) = 3.68 + 0.281 \ln(M)$, которое в нашем случае имеет высокую статистическую значимость ($p < 0.0001$) оцениваемого параметра β_2 при стандартном отклонении для остатков $s = 0.047$ и коэффициенте детерминации $R^2 = 0.876$.

Альтернативный путь заключается в построении и анализе модели байесовской регрессии: $Y = \beta_1 + \beta_2 X + \varepsilon$, $\varepsilon \sim N(0, \sigma^2)$. Для оцениваемых параметров β_1, β_2 и дисперсии остатков σ^2 формулируются априорные представления, которые выражаются плотностью вероятности их совместного распределения (β_1, β_2, σ^2). По результатам наблюдений, заключающихся в выборке значений X и Y, эти представления корректируются, т.е. с использованием теоремы Байеса ищется совместное апостериорное распределение ненаблюдаемых параметров при заданных данных пропорционально априорному распределению и правдоподобию: $p(\beta_1, \beta_2, \sigma^2|Y,X) \propto p(Y|X, \beta_1, \beta_2, \sigma^2)p(\beta_1, \beta_2|\sigma^2)p(\sigma^2)$.

Все эти вычисления мы можем выполнить с использованием обыкновенной версии алгоритма Метрополиса-Гастингса. Подробно комментируемый скрипт на языке статистической среды R для решения подобной задачи представлен Ф. Гартигом (Hartig) в его блоге, содержащем также и другие прекрасные материалы общего характера по экологии, эволюции и биогеографии. Мы предоставим возможность заинтересованному читателю выполнить такой расчет самостоятельно, а сами для прогона модели используем программную реализацию байесовских алгоритмов в программе JAGS.

Байесовская парадигма представлена в функциях многих пакетов статистической среды R. Однако наиболее привлекательной возможностью является использование вычислительной мощности бесплатных специализированных продуктов, которые сами по себе являются самостоятельными программными комплексами:

° проекта BUGS (акроним от Bayesian inference Using Gibbs Sampling – байесовская статистика с использованием генератора Гиббса) в реализациях WinBUGS (Лондонский императорский колледж) и OpenBUGS (Университет в Хельсинки) и

° программы JAGS (т.е. ″еще один генератор Гиббса″ – Just Another Gibbs Sampler), разработанной Международным агентством изучения рака.

Это – достаточно сложные программы, на вход которых подается описание модели, сделанное по определенным формальным правилам, а на выходе выдается цепь Маркова, сходящаяся к апостериорному распределению. Однако с помощью этих программ нельзя рисовать графики и неудобно осуществлять предварительную и завершающую обработку данных. Поэтому, несомненно, более удобно запускать их из среды R, для чего разработана следующая технология: а) пакеты типа rbugs, rjags и др. осуществляют интерфейс с установленными версиями BUGS или JAGS, т.е. специфицируют модель и задают начальные значения, после чего запускают функции внешних модулей этих программ; б) пакет coda принимает из внешней среды и обрабатывает синтезированную ″сырую″ цепь MCMC; в) функции пакетов типа ggmcmc осуществляют комплексную диагностику и графическую интерпретацию результатов.

Вернемся, однако, к ящерице прыткой. Зададим имитационной программе JAGS вид модели $Y \sim N(\beta_1, \beta_2 X, 1/\sigma^2)$, начальные (априорные) значения $\beta_1 = \beta_2 \sim N(0, 10^{-6})$ и желаемую длину цепи. Далее JAGS сам осуществляет выбор конкретного алгоритма имитации (генератор Гиббса, алгоритм "случайного блуждания" Метрополиса или какой-то другой специализированный выборочный процесс). Возвращаемая цепь (см. рис. 7.29а) содержит по 100000 случайных апостериорных значений для каждого параметра (β_1, β_2, σ^2), с использованием которых можно построить гистограмму, графики ядерной плотности (см. рис. 7.29б-г), автокорреляции или скользящей средней, оценить интервалы высокой плотности HPD и т.д. Нетрудно заметить, что для нашего случая байесовские оценки параметров в виде математического ожидания случайной величины практически не отличаются от оценок обычной регрессии, сделанных методом наименьших квадратов: $\beta_1 = 3.69$, $\beta_2 = 0.282$, $\sigma^2 = 0.0023$, $s = 0.048$.

Некоторые специализированные пакеты статистической среды R включают собственные функции генерации выборок из апостериорного распределения параметров рассчитываемых моделей с использованием методов Монте-Карло для марковских цепей. В разделе 4.3 мы рассматривали процедуру построения обобщенной линейной модели со смешанными эффектами (LMEM), реализованную с использованием функции lmer(…) из пакета lme4. Воспользуемся этой моделью для нашего третьего примера, ставящего задачу ответить на вопрос: является ли однородным пространственное распределение биомассы зоопланктона по акватории Куйбышевского водохранилища.

По исходным данным, связанным с этим примером, нами уже была рассмотрена (см. раздел 4.2) модель дисперсионного анализа с фиксированными эффектами, которые соответствовали трем изучаемым факторам:

$$Y = \mu + \text{STAN} + \text{MONTH} + \text{YEAR} + \{\text{комбинаторные эффекты}\} + \varepsilon,$$

где STAN – географическая изменчивость биомассы зоопланктона относительно шести станций наблюдения, расположенных в разных частях акватории; MONTH – каждый из 6 месяцев вегетативного периода, в течение которого велись ежегодные наблюдения; YEAR – многолетний тренд, рассматриваемый в контексте трех характерных периодов в истории водохранилища.

Если рассматривать оба показателя YEAR и MONTH временной динамики случайными факторами, оказывающими влияние на независимость повторностей, но не относящимися к сути решаемой задачи, то их вклад может быть представлен в составе модели со смешанными эффектами величинами вариаций S(MONTH) и S(YEAR):

$$Y = \mu + \text{STAN} + S(\text{MONTH}) + S(\text{YEAR}) + \varepsilon.$$

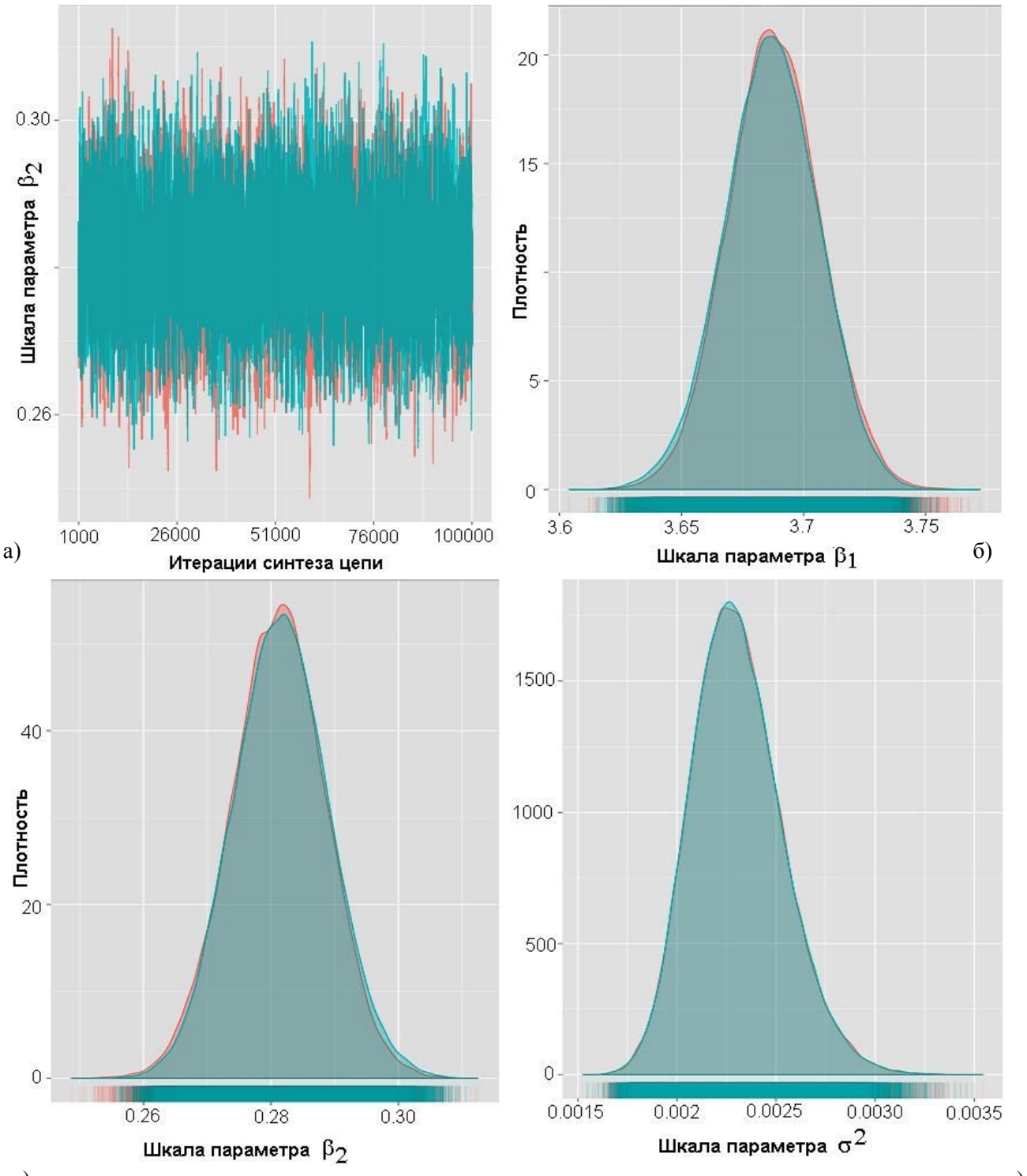


Рис. 7.29. График динамики значений марковской цепи (а) и распределения плотности апостериорных вероятностей параметров β_1, β_2 и σ^2 модели регрессии массы ящериц на длину тела (б-г)

Чтобы оценить статистическую значимость коэффициентов модели, связанных с уровнями фиксированного фактора (т.е. с местоположениями станций наблюдений), выполним генерацию 10000 выборок из апостериорного распределения этих параметров с использованием функции mcmcsamp(…). Графики распределения плотности вероятности значений коэффициентов для некоторых станций, представленные на рис. 7.30, дают нам полное представление о статистическом характере степени влияния этих уровней. Если задаться критической вероятностью доверия (например, 1 - α = 0.95), то можно оценить граничные значения интервала высокой апостериорной плотности HPD, т.е. 95% из 10000

значений марковской цепи будут находиться в пределах этого интервала – см. табл. 7.3. На основании этих расчетов можно предположить, что как минимум на трех станциях из 6 биомасса зоопланктона значимо отличается от средней величины для всего водохранилища.

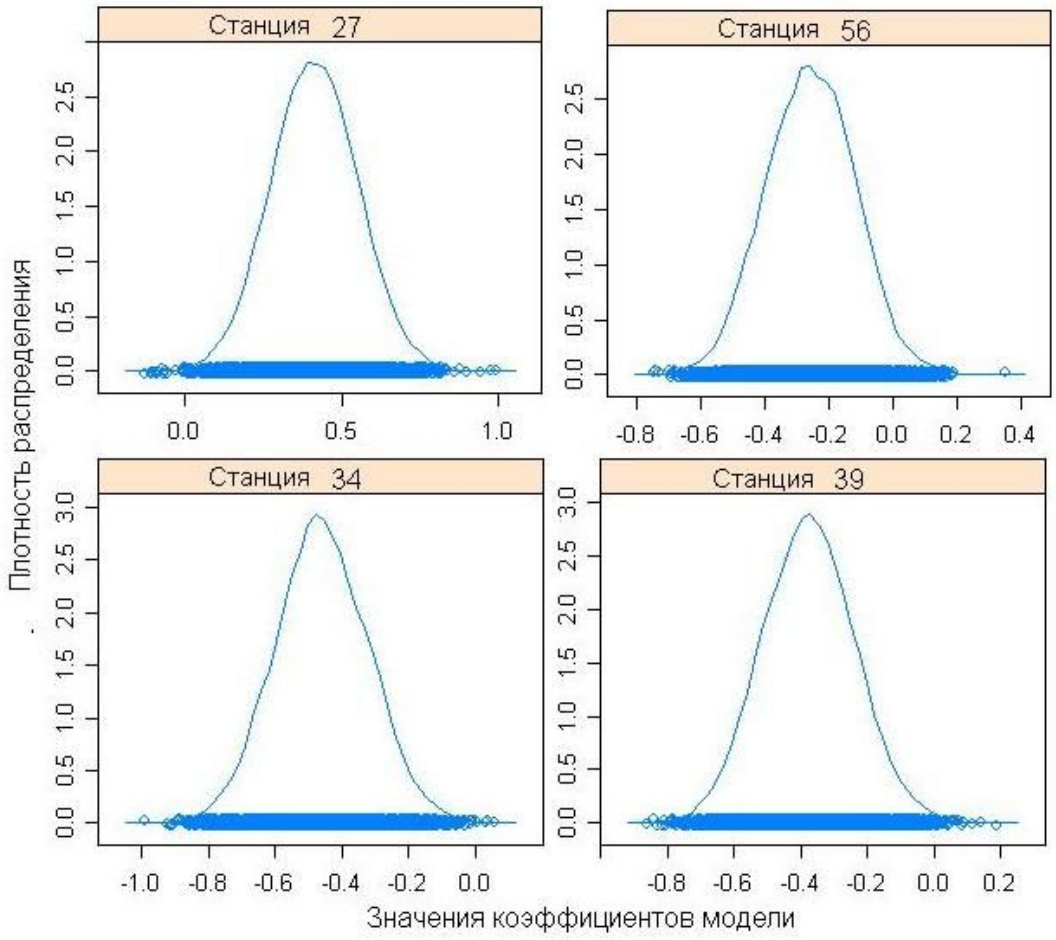


Рис. 7.30. Распределение плотности апостериорных вероятностей значений коэффициентов при фиксированном факторе (№ станции наблюдения) модели со смешанными эффектами для оценки изменчивости биомассы зоопланктона в Куйбышевском водохранилище

Таблица 7.3. Коэффициенты смешанной модели дисперсионного анализа вариации плотности зоопланктона на различных станциях наблюдения Куйбышевского водохранилища; HPD – диапазон высокой апостериорной плотности с вероятностью 0.95

Градации (номер станции)	Коэффициенты модели	*t*-значение	Интервалы HPD	
			нижний	верхний
20	-0.048	-0.35	-0.317	0.221
27	**0.416**	3.02	0.146	0.683
34	**-0.458**	-3.32	-0.735	-0.191
39	**-0.374**	-2.71	-0.646	-0.109
56	-0.256	-1.86	-0.514	0.019

К разделу 7.7:

```
#  ------------ Пример 1  -  Оценка вероятности дождя в норвежских фиордах
r=scan("blindern_tersklet.txt") ; k1=k2=n1=n2=k=0 ; n=length(r) ; k=sum(r)
for(i in 2:n) {
   if(r[i-1]==1)   {  # Из 596 дней с дождем в 202 случаях он продолжался на следующий день
       n1 = n1 + 1 ;   k1 = k1 + r[i]   }
   if(r[i-1]==0)    {   # Из 3055 дней без дождя в 393 случаях он начинался на следующий день
       n2 = n2 + 1 ;   k2 = k2 + r[i]   }
}
```

```
# Делим диапазон полной вероятности [0, 1] на 400 частей для построения функции распределения
p = p.m = p1 = p2 = seq(0.0025,1-0.0025,0.0025)
# Устанавливаем равные априорные вероятности
prior.p=prior.p1=prior.p2=rep(1/length(p),length(p))
# Рассчитываем правдоподобие параметров обоих моделей в логарифмической форме
lik.p=log(p)*k+log(1-p)*(n-k)
lik.p1=log(p1)*k1+log(1-p1)*(n1-k1) ;  lik.p2=log(p2)*k2+log(1-p2)*(n2-k2)
# Выполняем нормировку вероятностей на max.p и оценку апостериорных вероятностей
max.p=max(lik.p) ; lik.p=lik.p-max.p ; lik.p=exp(lik.p)
post.p=lik.p*prior.p ; post.p=post.p/sum(post.p)   #  Используем формулу Байеса
lik1=sum(lik.p*prior.p)  # Общее правдоподобие к данным для модели 1
# Аналогичные действия выполняем для модели 2
max.p1=max(lik.p1) ; lik.p1=lik.p1-max.p1 ; lik.p1=exp(lik.p1) ; post.p1=lik.p1*prior.p1
post.p1=post.p1/sum(post.p1)
max.p2=max(lik.p2) ; lik.p2=lik.p2-max.p2 ; lik.p2=exp(lik.p2) ; post.p2=lik.p2*prior.p2
post.p2=post.p2/sum(post.p2)
lik2=0 ; for(i in 1:length(p1))
              for(j in 1:length(p2))
    lik2=lik2+(p2[j]/(1-p1[i]+p2[j]))^r[1]*(1-p2[j]/(1-p1[i]+p2[j]))^(1-r[1])
                                              *lik.p1[i]*lik.p2[j]*prior.p1[i]*prior.p2[j]
B = lik1/lik2*exp(max.p-max.p1-max.p2)   #  Вычисляем байесовский фактор
par(mfrow=c(2,1))  #  Вывод графиков распределения плотности вероятностей
plot(p1, post.p1, xlim=c(0,0.42), type="h", xlab="p1", ylab="Pr(p1|D)", lwd=3)
plot(p2, post.p2, xlim=c(0,0.42), type="h", xlab="p2", ylab="Pr(p2|D)", lwd=3)
# Строим распределение разностей (p1 – p2)
diff.p = seq(-1+0.0025,1-0.0025,0.0025) ; post.diff.p=rep(0,length(diff.p))
for(i in 1:length(p1))
  for(j in 1:length(p2))
    {      index=i-j+length(p1)
      post.diff.p[index]=post.diff.p[index]+post.p1[i]*post.p2[j]  }
plot(diff.p,post.diff.p, xlim=c(0.1,0.3), type="h", xlab="p1-p2", ylab="Pr(p1-p2|D)", lwd=3)
sum(diff.p*post.diff.p)  ; sum(post.diff.p[diff.p<0])
#  ---------------  Пример 2 - Зависимость длины ящериц от массы тела
Z <- read.delim("Zootoca.txt")
# Строим классическую регрессионную модель аллометрии Huxley
y <- log(Z$svl) ; x <- log(Z$bm); lmod <- lm(y~x) ; summary(lmod)
plot(x,y, xlab='масса тела ln(), г',ylab='длина тела ln(), мм')
matplot(x,predict(lmod, interval="confidence"),type='l',lty=c(1,2,2), col=4, add=T)
# Строим модель байесовской регрессии
set.seed(3) ; n.obs <- length(x) ;  chain.len <- 100000  # количество наблюдений и длина цепи
# Выводим в файл описание модели и инициализацию априорных значений
write("model {
for (i in 1:n) { y[i] ~ dnorm(beta1+beta2*x[i],1/sigma2)}
  beta1 ~ dnorm(0,0.000001)
  beta2 ~ dnorm(0,0.000001)
  sigma2 ~ dunif(0,10^6) }
  }", "simple_regression.jags ")
library(rjags)  #  Выполняем интерфейс с JAGS-3.0 и передаем ему ссылку на файл с описаниями
model <- jags.model("simple_regression.jags",
                            data=list('y'=y,'x'=x,'n'=n.obs),n.chains=2,n.adapt=1000)
# Получаем файл с результатами
result <- coda.samples(model, variable.names=c("beta1","beta2","sigma2"),n.iter= chain.len)
# Получение статистик цепи и построение различных графиков
library(ggmcmc) ; df <- ggs(result)
summary(subset(df,Parameter=="beta2" & Chain==2, select=value))
ggs_density(df) ; ggmcmc(df, file = " Zootoca model.pdf", param.page = 1)
#  Пример 3  -  Пространственная неоднородность биомассы зоопланктона в Куйб. водохранилище
load("BSuzA.RData"); str(BSuzA) ; summary(aov(Level~Stan,BSuzA)) # Не учтен фактор  времени
# Многолетняя и сезонная изменчивости учитываются в случайных факторах смешанной модели
library(lme4) ; mod_s <- lmer(Level~Stan+(1|Year)+ (1|Month),BSuzA ) ;anova(mod_s)
# Формируем графики апостериорного распределения коэффициентов и находим HPD
cfi<-mcmcsamp(mod_s,n = 10000,saveb=T) ; HPDinterval(cfi) ; densityplot(cfi) # 10000 итераций
```

ЗАКЛЮЧЕНИЕ

Применение многомерных статистических методов и алгоритмов распознавания образов в экологических исследованиях имеет давнюю историю (Розенберг, 1977, 1980, 1981, 1984). Принципиальная сложность взаимодействия природных систем с факторами окружающей среды еще задолго до появления первых работ по бутстепу предопределило разработку комьютерно-интенсивных методов обработки данных, таких как расчет меры диссонанса исходной и рандомизированной матриц связи (Розенберг, 1975), комбинаторные способы оценки устойчивости ценозов (Розенберг и др., 1980), алгоритм «модельного штурма» (Брусиловский, Розенберг, 1983) и т. д. Настоящая монография подводит своеобразный промежуточный итог долгому пути в этом направлении.

Ограничившись подробным рассмотрением методов ресамплинга и других алгоритмов из семейства Монте-Карло, авторы ни в коем случае не стремились противопоставить их традиционным методикам статистического анализа (регрессионному анализу, различным «старым» моделям прогнозирования временных рядов, кластерному анализу и т. д.). Эти методы, основанные на серьезной теоретической платформе и выдержавшие проверку десятилетиями, также нужно всемерно изучать и использовать. В качестве примера гармонического сочетания «старого и нового» можно привести программное обеспечение эколого-информационной системы Волжского бассейна (REGION-VOLGABAS – Розенберг, 2009), где, наряду с общепринятыми методами статистики, для прогнозирования сценариев возможного развития региона используются различные модели самоорганизации (эволюционное и нейросетевое моделирование, метод группового учета аргументов, карты Кохонена и др.).

С другой стороны, у ученых «пришло понимание субъективности образа экологического мира: он перестал быть понятным и объяснимым, а его познание перешло из стадии созерцательной неподвижной гармонии к потоку нескончаемых изменений» (Розенберг, Смелянский, 1997). И здесь методы Монте-Карло – только одно из возможных направлений развития. Современный взгляд на биотические сообщества открывает широкие пути применения для анализа их структуры и многих других достижений физики и математики, включая синергетику, кибернетику, теорию сложности, концепцию самоорганизованной критичности. Например, одним из таких направлений является использование фрактальной методологии при объяснении инвариантных характеристик структуры экосистем, эффективность которой подробно и последовательно обсуждается в недавно вышедшей монографии (Гелашвили и др., 2013).

Поэтому авторам остается только пожелать молодым исследователям помнить слова И. Пригожина (2000, с. 14): «На наших глазах рождается наука, не ограничиваемая более идеализированными и упрощенными ситуациями, а отражающая всю сложность реального мира…».

СПИСОК ЛИТЕРАТУРНЫХ ИСТОЧНИКОВ

Айвазян С.А. Байесовский подход в эконометрическом анализе // Прикладная эконометрика. 2008. Т.9, №1. С. 93–130.

Айвазян С.А., Буштабер В.М., Енюков И.С., Мешалкин Л.Д. Прикладная статистика. Классификация и снижение размерностей. М.: Финансы и статистика, 1989. 607 с.

Айвазян С.А., Мхитарян В.С. Прикладная статистика и основы эконометрии. М.: ЮНИТИ, 1998. 1022 с.

Алгоритмы и программы восстановления зависимостей. М.: Наука, 1984. 816 с.

Анатольев С. Основы бутстрапирования // Квантиль. 2007. №3. С. 1-12.

Анатольев С. Непараметрическая регрессия // Квантиль. 2009. №7. С. 37-52.

Афифи А., Эйзен С. Статистический анализ: Подход с использованием ЭВМ. М.: Мир, 1982. 488 с.

Бидюк П.И., Павлов В.В., Борисевич А.С. и др. Оценивание регрессионных моделей с помощью метода Монте-Карло для марковских цепей // Кибернетика и вычисл. техника. 2009. Вып. 156. С. 40-57.

Бир С. Кибернетика и управление производством. М.: Наука, 1963. 276 с.

Бокс Дж., Дженкинс Г. Анализ временных рядов. Прогноз и управление. М.: Мир, 1974. Вып. 1. 406 с.

Вапник В.Н., Червоненкис А.Я. Теория распознавания образов. М.: Наука, 1974. 487 с.

Воробейчик Е.Л. О некоторых индексах ширины и перекрывания экологических ниш // Журн. общ. биологии. 1993. Т. 54, № 6. С. 706-712.

Гайдышев И. Анализ и обработка данных: специальный справочник. СПб: Питер, 2001. 752 с.

Гелашвили Д.Б., Иудин Д.И., Розенберг Г.С. и др. Фракталы и мультифракталы в биоэкологии. Нижний Новгород: Изд-во Нижегородского госуниверситета, 2013. 370 с.

Горбань А.Н., Дунин-Барковский В.Л., Миркес Е.М. и др. Нейроинформатика. Новосибирск: Наука. Сиб. предприятие РАН, 1998. 296 с.

Грабарник П.Я. Статистические методы в пространственной экологии: новый взгляд на типичные задачи анализа экологических данных // Тезисы II Нац. научн. конф. «Математическое моделирование в экологии». Пущино: ИФХиБПП РАН, 2011. С. 80-82.

Грабарник П.Я., Комаров А.С. Статистический анализ пространственных структур. Методы, использующие расстояния между точками. Пущино: НЦ БИ АН СССР, 1980. 48 с.

Дрейпер Н., Смит Г. Прикладной регрессионный анализ. М.: Финансы и статистика. Кн. 1, 1986. 366 с. Кн. 2, 1987. 352 с.

Дэйвисон М. Многомерное шкалирование. Методы наглядного представления данных. М.: Финансы и статистика, 1988. 348с.

Дюран Б., Оделл П. Кластерный анализ. М.: Статистика, 1977. 128 с.

Елисеева Л.И., Рукавишников В.О. Группировка, корреляция, распознавание образов. М.: Статистика, 1977. 144 с.

Ефимов В.М., Галактионов Ю.К., Шушпанова Н.Ф. Анализ и прогноз временных рядов методом главных компонент. М.: Наука, 1988. 70 с.

Ефимов В.М., Ковалева В.Ю. Многомерный анализ биологических данных: учебное пособие. Горно-Алтайск: РИО-ГАГУ, 2007. 75 с.

Загоруйко Н.Г. Прикладные методы анализа данных и знаний. Новосибирск: ИМ СО РАН, 1999. 270 с.

Закс Л. Статистическое оценивание. М.: Статистика, 1976. 598 с.

Зарядов И.С. Введение в статистический пакет R: типы переменных, структуры данных, чтение и запись информации, графика. М.: Издательство Российского университета дружбы народов, 2010а. 207 с.

Зарядов И.С. Статистический пакет R: теория вероятностей и математическая статистика. М.: Издательство Российского университета дружбы народов, 2010б. 141 с.

Зиновьев А.Ю. Визуализация многомерных данных. Красноярск: КГТУ, 2000. 168 с.

Кендалл М., Стьюарт А. Статистические выводы и связи. М.: Наука, 1973. 899 с.

Кендалл М., Стьюарт А. Многомерный статистический анализ и временные ряды. М.: Наука, 1976. 736 с.

Классификация и кластер. / Под ред. Дж. Вэн-Райзина. М.: Мир, 1980. 390 с.

Кобзарь А.И. Прикладная математическая статистика. Для инженеров и научных работников. М.: Физматлит, 2006. 816 с.

Кохонен Т. Ассоциативные запоминающие устройства. М.: Мир, 1982. 383 с.

Лбов Г.С. Методы обработки разнотипных экспериментальных данных. Новосибирск: Наука, 1981. 160 с.

Левич А.П. Структура экологических сообществ. М.: Изд-во МГУ, 1980. 182 с.

Миркин Б.М., Розенберг Г.С. Фитоценология. Принципы и методы. М.: Наука, 1978. 212 с.

Монтгомери Д.К. Планирование эксперимента и анализ данных. Л.: Судостроение, 1980. 384 с.

Мостеллер Ф., Тьюки Дж. Анализ данных и регрессия. М: Финансы и статистика, 1982. Вып. 1. 320 с.

Орлов А И. Прикладная статистика. М.: Экзамен, 2007. 671 с. URL: http://orlovs.pp.ru

Орлов А И. Эконометрика. М.: Экзамен, 2002. 576 с. URL: http://orlovs.pp.ru

Павлинов И.Я., Микешина И.Г. Принципы и методы геометрической морфометрии // Журн. общей биологии. 2002. Т. 63, № 6. С. 473-493.

Пригожин И.Р. Конец определенности. Время, хаос и новые законы природы. Ижевск: НИЦ «Регулярная и хаотическая динамика», 2000. 208 с.

Пузаченко Ю.Г. Математические методы в экологических и географических исследованиях. М: Академия, 2004. 416 с.

Растригин Л.А., Эренштейн Р.Х. Метод коллективного распознавания. М.: Энергоатомиздат, 1981. 80 с.

Розенберг Г.С. Обзор методов статистической геоботаники. 3. Методы автоматической классификации. М., 1977. 38 с. Деп. в ВИНИТИ 11.04.1977. № 1321-77.

Розенберг Г.С. Вероятностный подход к изучению временной структуры растительного покрова // Журн. общ. биол. 1980. Т. 41, № 3. С. 372-385.

Розенберг Г.С. Сравнение различных методов экологического прогнозирования. Прогноз динамики экосистем. // Экология. 1981. № 1. С. 12-18.

Розенберг Г.С. Модели в фитоценологии. М.: Наука, 1984. 240 с.

Розенберг Г.С. К методике использования теории распознавания образов в фитоиндикационных исследованиях // Статистические методы классификации растительности и оценки ее связи со средой. Уфа: БФАН СССР, 1975. С. 5-37

Розенберг Г.С., Назарова З.М., Миркин Б.М. О способе оценки устойчивости фитоценозов // Бюлл. МОИП. Отдел биол. 1980. Т. 85, вып. 1. С. 129-131.

Брусиловский П.М., Розенберг Г.С. Модельный штурм при исследовании экологических систем // Журн. общ. биол. 1983. Т. 44. № 2. С. 254-262.

Розенберг Г.С. Волжский бассейн: на пути к устойчивому развитию. Тольятти: ИЭВБ РАН; Кассандра, 2009. 477 с.

Розенберг Г.С., Смелянский И.Э. Экологический маятник (смена парадигм в современной экологии) // Журн. общ. биол. 1997. Т. 58. № 4. С. 5-19.

Розенберг Г.С., Шитиков В.К., Брусиловский П.М. Экологическое прогнозирование (Функциональные предикторы временных рядов). Тольятти: ИЭВБ РАН, 1994. 185 с. URL: http://www.ievbras.ru/ecostat/Kiril/Download/ Mevr.pdf

Савельев А.А., Мухарамова С.С., Пилюгин А.Г. и др. Геостатистический анализ данных в экологии и природопользовании (с применением пакета R). Казань: Казанский университет, 2012. 120 с. URL: http://gis-lab.info/docs/saveliev2012-geostat.pdf

Слуцкий Е.Е. Сложение случайных причин как источник циклических процессов // Вопросы конъюнктуры. М.: Финиздат НКФ, 1927. Т. III. Вып. 1. С. 37-61.

Советы молодому ученому: методическое пособие для студентов, аспирантов, младших научных сотрудников и, может быть, не только для них. Екатеринбург: ИЭРиЖ УрО РАН, 2011. 122 с.

Стрижов В.В., Крымова Е.А. Методы выбора регрессионных моделей. М.: ВЦ РАН, 2010. 60 с.

Тейл Г. Экономические прогнозы и принятие решений. М.: Статистика, 1971. 488 с.

Тутубалин В.Н., Барабашева Ю.М., Григорян А.А. и др. Математическое моделирование в экологии (Историко-методологический анализ). М.: Языки русской культуры, 1999. 208 с.

Уоссермен Ф. Нейрокомпьютерная техника. М.: Мир, 1992. 184 с.

Хайтун С.Д. Негауссовость социальных явлений // Социологические исследования. 1983. № 1. С. 144-152.

Хромов-Борисов Н.Н. Синдром статистической снисходительности или значение и назначение *p*-значения // Телеконференция по медицине, биологии и экологии, 2011. №4. URL: http://tele-conf.ru/aktualnyie-problemyi-tehnologicheskih-izyiskaniy/3.html .

Цейтлин Н.А. Из опыта аналитического статистика. М.: Солар, 2007. 906 с. URL: http://www.cubematrix.com/oldsite/anlagen/as.pdf

Цейтлин Н.А. Статистический подход к оцениванию знаний учащихся // Комп'ютерне моделювання в хімії, технологіях і системах сталого розвитку. Київ, Рубіжне: НТУУ «КПІ», 2012. С. 254-256.

Цыплаков А. Мини-словарь англоязычных эконометрических терминов, часть 2. // Квантиль. 2008. №5. С. 41-48.

Чижикова Н.А. Некоторые закономерности формирования дискретных структур в континууме растительного покрова. Дис. канд. биол. наук. Казанский государственный университет, 2008. 141 с.

Шипунов А.Б., Балдин Е.М., Волкова П.А. и др. Наглядная статистика. Используем R! М.: ДМК Пресс, 2012. 298 с.

Шитиков В.К. Использование рандомизации и бутстрепа при обработке результатов экологических наблюдений // Принципы экологии (научный электронный журнал). 2012, № 1. С. 4 – 24. URL: http://ecopri.ru/journal/atricle.php?id=481

Шитиков В.К., Зинченко Т.Д. Комплексные критерии экологического состояния водных объектов: экспертный и статистический подход // Количественные методы экологии и гидробиологии (Сборник научных трудов, посвященный памяти А.И. Баканова) / Отв. ред. чл.-корр. Г.С. Розенберг. Тольятти: СамНЦ РАН, 2005. С. 134-148. URL: http://www.ievbras.ru/ecostat/Kiril/Download/Meba.pdf .

Шитиков В.К., Зинченко Т.Д. Анализ статистических закономерностей организации видовой структуры донных речных сообществ // Журнал общей биологии. 2011. Т. 72, № 5. С. 355–368.

Шитиков В.К., Зинченко Т.Д. Изменение таксономического и функционального разнообразия сообществ макрозообентоса по продольному градиенту рек // Успехи современной биологии. 2013. Т. 133, № 6. С. 566-577.

Шитиков В.К., Зинченко Т.Д., Абросимова Э.В. Непараметрические методы сравнительной оценки видового разнообразия речных сообществ макрозообентоса // Журнал общей биологии. 2010. Т. 71, № 3. С. 263-274.

Шитиков В.К., Зинченко Т.Д., Абросимова Э.В. Статистический анализ результатов многомерной ординации на примере донных речных сообществ // Экология. 2012. № 1. С. 1-6.

Шитиков В.К., Зинченко Т.Д., Розенберг Г.С. Макроэкология речных сообществ: концепции, методы, модели. Тольятти: СамНЦ РАН, Кассандра, 2012. 257 с. URL: http://www.ievbras.ru/ecostat/Kiril/Download/Maec.pdf .

Шитиков В.К., Розенберг Г.С., Зинченко Т.Д. Количественная гидроэкология: методы, критерии, решения: В 2-х кн. – М.: Наука, 2005.
Кн. 1. – 281 с. URL: http://www.ievbras.ru/ecostat/Kiril/Download/Meke1.pdf;
Кн. 2. – 337 с. URL: http://www.ievbras.ru/ecostat/Kiril/Download/Meke2.pdf.

Шитиков В.К., Розенберг Г.С., Крамаренко С.С., Якимов В.Н. Современные подходы к статистическому анализу экспериментальных данных // Проблемы экологического эксперимента (Планирование и анализ наблюдений). Тольятти: СамНЦ РАН, Кассандра, 2008. С. 212-250.
URL: http://www.ievbras.ru/ecostat/Kiril/Download/Mepe.pdf.

Эфрон Б. Нетрадиционные методы многомерного статистического анализа. М.: Финансы и статистика, 1988. 263 с.

Albert J., Rizzo M. R by Example. N.Y.: Wiley, 2004. 359 p.

Amaral G.J., Dryden I.L., Wood A.T. Pivotal bootstrap methods for *k*-sample problems in directional statistics and shape analysis // Journal of the American Statistical Association. 2007. V. 102. P. 695-707.

Anderson M.J. A new method for non-parametric multivariate analysis of variance // Austral. Ecology. 2001. V. 26. P. 32-46.

Anderson M.J. Distance-based tests for homogeneity of multivariate dispersions // Biometrics. 2006. V. 62. P. 245-253.

Anderson M.J., ter Braak C.J.F. Permutation tests for multi-factorial analysis of variance // Journal of Statistical Computation and Simulation. 2003. V. 73. P. 85-113.

Andrieu C., de Freitas N., Doucet A. et al. An introduction to MCMC for machine learning // Mach. Learn. 2003. V. 50. P. 5–43.

Baddeley A. Analyzing spatial point patterns in R. Workshop notes. Version 4.1. CSIRO online technical publication. 2010. 232 p. URL: www.csiro.au/resources/pf16h.html

Barnard G.A. Discussion of paper by M.S. Bartlett // J. R. Stat. Soc. 1993. V. 25. 294 p.

Bayes T. An essay towards solving a problem in the doctrine of chances. Philosophical Transactions of the Royal Society. 1763. P. 330-418. Reprinted in: Biometrica. 1958. № 45. P. 293-315.

Bezdek J.C. Partition structures: a tutorial // Analysis of Fuzzy Information. Florida: CRC Press, Boca Raton, 1987. V. 3. P. 81-107.

Bivand R., Pebesma E.J., Gomez-Rubio V. Applied spatial data analysis with R. N.Y.: Springer, 2008. 374 p.

Blanchet F.G., Legendre P., Borcard D. Forward selection of explanatory variables. // Ecology. 2008. V. 89. P. 2623-2632.

Bolker B. Ecological models and Data in R. Princeton: University Press, 2007. 507 p.

Bookstein F.L. Morphometric tools for landmark data: geometry and biology. Cambridge: Cambridge Univ. Press, 1991. 198 p.

Borcard D., Gillet F., Legendre P. Numerical Ecology with R. N.Y.: Springer, 2011. 306 p.

Box G., Cox D.R. An analysis of Transformation // Journal of Royal Statistical Society B. 1964. V. 26. P. 211-243.

Boyce R.L, Ellison P.C Choosing the best similarity index when performing fuzzy set ordination on binary data // Journal of Vegetation Science. 2001. V. 12. P. 711-720.

Breiman L. Random forests // Machine Learning. 2001. V. 45, № 1. P. 5-32.

Breiman L., Friedman J.H., Olshen R.A. et al. Classification and Regression Trees. Belmont (CA): Wadsworth Int. Group, 1984. 368 p.

Burnham K. P., Anderson D. R. Model selection and multimodel inference: a practical information-theoretic approach. N.Y.: Springer-Verlag, 2002. 496 p.

Chao A., Chazdon R.L., Colwell R.K. et al. A new statistical approach for assessing similarity of species composition with incidence and abundance data // Ecol. Letters. 2005. V. 8. P. 148-159.

Chao A., Shen T.J. Nonparametric estimation of Shannon`s index of diversity when there are unseen species in sample // Environmental and Ecol. Statist. 2003. V. 10. P. 429-443.

Chaves L.F. An Entomologist guide to demystify Pseudoreplication: data analysis of field studies with design constraints // Journal of medical entomology. 2010. V. 47, № 3. P. 291-298.

Chernick M.R. Bootstrap methods, a practitioner's guide. Hoboken (NJ): Wiley, 1999. 369 p.

Chernick M.R., Fritis R. Introductory biostatistics for the health sciences: modern applications including bootstrap. Hoboken (NJ): Wiley, 2003. 406 p.

Clarke K.R. Non-parametric multivariate analyses of changes in community structure // Austral. J. Ecol. 1993. V. 18. P. 117-143.

Clarke K.R., Warwick R.M. A taxonomic distinctness index and its statistical properties // J. Appl. Ecol. 1998. V. 35. P. 523-531.

Claude J. Morphometrics with R. Use R! series. N.Y.: Springer, 2008. 316 p.

Clauset A., Shalizi C. R., Newman M. E. Power-law distributions in empirical data // SIAM Review, 2009. V. 51. P. 661-703.

Cleveland W.S. Robust locally weighted regression and smoothing scatterplots // Journal of the American Statistical Association. 1979. V. 74, № 368. P. 829-836.

Cowpertwait P.S., Metcalfe A.V. Introductory Time Series with R. N.Y.: Springer, 2009. 251 p.

Crawley M.J. The R Book. London: Wiley & Sons Inc., 2007. 930 p.

Davison A.C., Hinkley D.V. Bootstrap methods and their application. Cambridge: Cambridge University Press, 2006. 592 p.

Davison R. A., Kuonen D. An Introduction to the Bootstrap with Applications in R // Statistical Computing and Statistical Graphics Newsletter. 2002. V. 13**, №** 1. P. 6-11.

De Cáceres M, Legendre P. Associations between species and groups of sites: indices and statistical inference // Ecology. 2009. V. 90, № 12. P. 3566-3574.

De'Ath G. Multivariate regression trees: a new technique for modeling species environment relationships // Ecology. 2002. V.83. P. 1105-1117.

Demidenko E. Mixed Models - Theory and Applications. Hoboken (NJ): Wiley-Interscience, 2004. 704 p.

Dickey D.A., Fuller W.A. Distribution of the estimators for autoregressive time series with a unit root // J. Amer. Statist. Ass. 1979. V. 74. P. 427-431.

Dryden I.L., Mardia K.V. Statistical shape analysis. N.Y.: Wiley, 1998. 347p.

Dufrene M., Legendre P. Species assemblages and indicator species: the need for a flexible asymmetrical approach // Ecol. Monogr. 1997. V.67, №3. P. 345-366.

Edgington E.S. Randomization tests. N.Y.:Marcel Dekker, 1995. 341 p.

Efron B. Computers and the theory of statistics: thinking the unthinkable // SIAM Review. 1979a. V. 21, № 4. P. 460-480.

Efron B. Bootstrep methods. Another look at the Jacknife // Ann. Statist. 1979б. № 7. P. 1-26.

Efron B., Tibshirani R.J. An introduction to the bootstrap. N.Y.: Chapman & Hall, 1993. 436 p.

Everitt B.S., Hothorn T. A Handbook Of Statistical Analyses Using R. N.Y.: Chapman and Hall, 2006. 304 p.

Everitt B.S., Howell D.C. Encyclopedia of Statistics in Behavioral Science. Chichester: Wiley & Sons Ltd., 2005. 2132 p.

Faraway J.J. Extending the linear model with R: generalized linear, mixed effects and nonparametric regression models. Boca Raton (FL): Chapman & Hall, 2006. 331 p.

Felsenstein J. Confidence limits on phylogenies: an approach using the bootstrap // Evolution. 1985.V.39. P. 783-791.

Fisher R.A. The arrangement of field experiments // Journal of the Ministry of Agriculture of GreatBritain. 1926. Vol. 33. P. 700-725.

Fisher R.A. The Design of Experiments. 8-th edit. N.Y.: Hafner Pub. Co., 1966. 248 p. (1-th edit. 1935).
Fox J. An R and S-Plus Companion to Applied Regression. London: Sage Publications Inc., 2002. 328 p.
Garnier E., Cortez J., Billes G. at al. Plant functional markers capture ecosystem properties during secondary succession // Ecology. 2004. V. 85. P. 2630-2637.
Goldberg D.E. Genetic algorithms in search, optimization and machine learning. Reading (MA): Addison Wesley, 1989. 432 p.
Good P. Introduction to Statistics Through Resampling Methods and R/S-Plus. N.Y.: Springer, 2005a. 315 p.
Good P. Permutation, parametric and bootstrap tests of hypotheses. N.Y.: Springer, 2005б. 315 p.
Good P. Resampling Methods: a practical guide to data analysis. N.Y.: Springer, 2006. 218 p.
Goodall C.R. Procrustes methods in the statistical analysis of shape // Journal of the Royal Statistical Society. 1991. Series B, V. 53. P. 285-339.
Goodman L.A., Kruskal W.H. Measures of association for cross classifications // Journal of the American Statistical Association. 1954. V. 49. P. 732-764.
Gordon A.D. Classification. London: Chapman and Hall/CRC, 1999. 248 p.
Gotelli N.J. Null model analysis of species co-occurrence patterns // Ecology. 2000. V. 81. P. 2606-2621.
Gotelli N.J., Graves G.R. Null Models in Ecology. Washington (DC): Smithonian Inst. Press, 1996. 368 p.
Gotelli N.J., Entsminger N.J. Swap algorithms in null model analysis // Ecology. 2003. V. 84. P. 532-535.
Gotelli N.J., McGill B.J. Null versus neutral models: what's the difference? // Ecography. 2006. V. 29. P. 793-800.
Guisan A., Thuillier W. Predicting species distribution: offering more than simple habitats models // Ecol. letters. 2005. V. 8. P. 993-1009.
Hartig F., Calabrese J.M., Reineking B. et al. Statistical inference for stochastic simulation models – theory and application // Ecol. Lett. 2011. V.14. P. 816-827.
Hastie T., Tibshirani R. Generalized Additive Models. London: Chapman & Hall, 1990. 335 p.
Hastie T., Tibshirani R., Friedman J. The Elements of Statistical Learning: Data Mining, Inference and Prediction. N.Y.: Springer-Verlag. 2009. 763 p
Helsel D. R., Hirsch R. Statistical methods in water resources. Techniques of Water Resources Investigations.U.S.: Geological Survey, 2002. Chap. A3, book 4. 522 p.
Henry M., Stevens H. A Primer of Ecology with R. N.Y.: Springer, 2009. 402 p.
Hothorn T, Hornik K, Zeileis A. Unbiased Recursive Partitioning: A Conditional Inference Framework // Journal of Computational and Graphical Statistics. 2006. V. 15, № 3. P. 651-674.
Howell D. C. Statistical Methods for Psychology. Wadsworth: Cengage Learning, 2009. 793 p.
Howell D. C. Fundamental Statistics for the Behavioral Sciences. Wadsworth: Cengage Learning, 2010. 676 p.
Huisman J., Olff H., Fresco L.F. A hierarchical set of models for species response analysis // J. Veg. Sci. 1993. V. 4. P. 37-46.
Hurlbert S.H. Pseudoreplication and the design of ecological field experiments // Ecological Monographs. 1984. V. 54. P. 187-211. Перевод см: Проблемы экологического эксперимента (Планирование и анализ наблюдений). Тольятти: СамНЦ РАН, Кассандра, 2008. С. 9-45. URL: http://www.ievbras.ru/ecostat/Kiril/Download/Mepe.pdf .
Hyndman R., Koehler A.B., Ord J.K. et al. Forecasting with exponential smoothing: the state space approach. Berlin: Springer-Verlag, 2008. 376 p.
Jongman R.H.G., ter Braak C.J.F., van Tongeren O.F.R. Data Analysis in Community and Landscape Ecology. Wageningen: Pudoc, 1987. 299 p. [Пер. с англ.: Джонгман Р.Г.Г.,

тер Браак С.Дж.Ф., ван Тонгерен О.Ф.Р. Анализ данных в экологии сообществ и ландшафтов. М.: РАСХН. 1999. 306 с.].

Jung K. Least Trimmed Squares Estimator in the Errors-in-Variables Model // Journal of Applied Statistics. 2007. V. 34, № 3. P. 331-338.

Kaufman L., Rousseeuw P.J. Finding Groups in Data. An Introduction to Cluster Analysis. Hoboken (NJ): Wiley & Sons Inc., 2005. P. 342.

Kempthorne O. Sampling inference, experimental inference and observation inference. // Sankhya: The Indian Journal of Statistics, Series B. 1979. V. 40, 3/4. P. 115-145.

Kempthorne O., Doerfler T.E. The Behaviour of Some Significance Tests under Experimental Randomization // Biometrika. 1969. V. 56. P. 231-248.

Kohonen T. Self-organized formation of topologically correct feature maps // Biological Cybernetics. 1982. № 43. P. 59-69.

Khromov-Borisov N.N., Henriques J. A.P. Good statistics practice (GSP) in genetic toxicology // Mutat. Res. 1998. V. 405, N. 1. P. 97-108.

Kuan C.-M., Hornik K. The generalized fluctuation test: A unifying view // Econometric Reviews. 1995. V. 14. P. 135-161.

Legendre P., Gallagher E. Ecologically meaningful transformations for ordination of species data // Oecologia. 2001. V. 129. P. 271-280.

Legendre P., Legendre L. Numerical Ecology. Amsterdam: Elsevier Sci. BV, 1998. 853 p.

Lepš J., de Bello F., Lavorel S. et al. Quantifying and interpreting functional diversity of natural communities: practical considerations matter // Preslia. 2006. V. 78. P. 481–501.

Link W.A., Baker R.J. Bayesian Inference: with Ecological Applications. London: Elsevier Ltd. Academic Press. 2009. 339 p.

Liu H., Ciannelli L., Decker M.B. et al. Nonparametric Threshold Model of Zero-Inflated Spatio-Temporal Data with Application to Shifts in Jellyfish Distribution // Journal of Agricultural, Biological and Environmental Statistics. 2011. V. 16. P. 185-201

Logan M. Biostatistical Design and Analysis Using R: A Practical Guide. Chichester: Wiley-Blackwell, 2010. 576 p.

Lunneborg C. E. Data analysis by resampling: Concepts and applications. Pacific Grove (CA): Duxbury, 2000. 568 p.

MacArthur R.H. On the relative abundance of bird species // Proc. Natl. Acad. Sci. USA. 1957. V. 45. P. 293-295.

Maindonald J., Braun W.J. Data Analysis and Graphics Using R. Cambridge: University Press, 2010. 525 p.

Manly B.F.J. Randomization, bootstrap and Monte Carlo methods in biology. London:Chapman & Hall, 2007. 445 p.

Mantel N. The detection of disease clustering and a generalized regression approach // Cancer Res. 1967. V. 27. P. 209-220.

Marcus L., Corti M., Loy A. et al. Advances in morphometrics. N.Y., London: Plenum Press, 1996. 587 p.

Mason N.W., Mouillot D., Lee W.G. et al. Functional richness, functional evenness and functional divergence: the primary components of functional diversity // Oikos. 2005. V. 111. P. 112-118.

McArdle B.H., Anderson M.J. Fitting multivariate models to community data: a comment on distance-based redundancy analysis // Ecology. 2001. V. 82, № 1. P. 290-297.

McCarthy M.A. Bayesian Methods for Ecology. Cambridge: University Press, 2007. 310 p.

McCullagh P., Nelder J.A. Generalized Linear Models. London: Chapman & Hall, 1989. 511 p.

McCune B., Grace J.B., Urban D.L. Analysis of Ecological Communities. Gleneden Beach (OR): MjM Software, 2002. 285 p.

Mielke P.W. Meteorological applications of permutation techniques based on distance functions // Handbook of Statistics. V. 4. Nonparametric Methods. Amsterdam: North-Holland, 1984. P. 813-830.

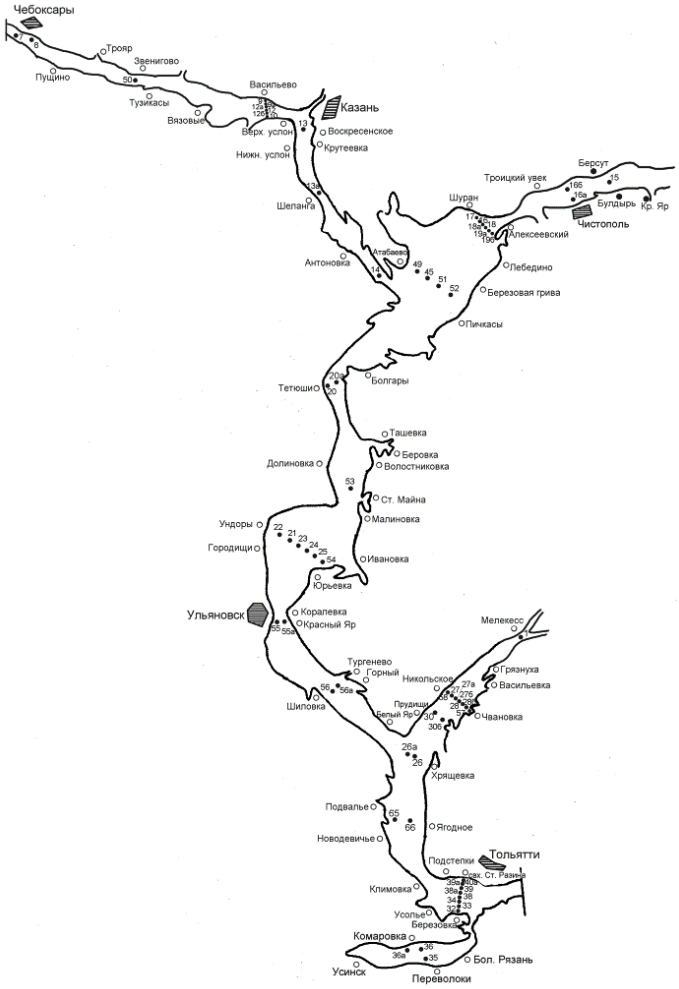
Mielke P.W., Berry K.J. Permutation Methods: A Distance Function Approach. N.Y.: Springer-Verlag, 2001. 357 p.
Miklós I., Podani J. Randomization of presence-absence matrices: comments and new algorithms // Ecology. 2004. V. 85. P. 86-92.
Miller A.J. Subset Selection in Regression. N.Y.: Chapman and Hall, 1990. 256 p.
Minchin P.R. An evaluation of the relative robustness of techniques for ecological ordination // Vegetatio. 1987. V. 67. P. 1167-1179.
Mooney C.Z., Duval R.D. Bootstrapping. A nonparametric approach to statistical inference. Sage (CA): University paper, 1993. 80 p.
Moore D. Bootstrap Methods and Permutation Tests // The Practice of Business Statistics. Ed. T. Hesterberg. N.Y.: Freeman & C^{o}, 2003. Cap. 14. 70 p.
Oberhofer W., Kmenta J. A General Procedure for Obtaining Maximum Likelihood Estimates in Generalized Regression Models // Econometrica, 1974. V. 42, № 3. P. 579-590.
Oksanen J., Minchin P.R. Continuum theory revisited: what shape are species responses along ecological gradients? // Ecol. Modelling. 2002. V. 157. P. 119-129.
Paradis E. Analysis of Phylogenetics and Evolution with R. N.Y.: Springer, 2009. 211 p.
Patefield W.M. Algorithm AS159. An efficient method of generating *r* x *c* tables with given row and column totals // Applied Statistics. 1981. V. 30. P. 91-97.
Perry J.A., Schaeffer D.J. The longitudinal distributions of or riverine benthos: a river discontinuum? // Hydrobiologia. 1987. V. 148. P. 257-268.
Petchey O.L., Gaston K.J. Dendrograms and measuring functional diversity // Oikos. 2007. V. 116. P. 1422-1426.
Petchey O.L., Gaston K.J. Functional diversity (FD), species richness, and community composition // Ecology Letters. 2002. V. 5. P. 402-411.
Pielou E.C. Ecological Diversity. N.Y.: Wiley, 1975. 165 p.
Pillar V.D. The bootstrapped ordination reexamined. // J. Vegetation Sci. 1999. V. 10. P. 895-902.
Pinheiro, J. C., Bates D.M. Mixed effects models in S and S-plus. N.Y.: Springer, 2000. 528 p.
Rao C.R. Diversity and dissimilarity coefficients: a unified approach // Theor. Popul. Biol. 1982. V. 21. P. 24-43.
Rao C.R. The use and interpretation of principal component analysis in applied research // Sankhya. 1964. Ser. A. V. 26. P. 329-358.
Revelle W., Rocklin T. Very Simple Structure: an alternative procedure for estimating the optimal number of interpretable factors // Multivariate Behavioral Research. 1979. V. 14. P. 403-414.
Ricotta C. Through the jungle of biological diversity // Acta Biotheor. 2005. V. 53. P. 29-38.
Ricotta C., Avena G.C. An information-theoretical measure of taxonomic diversity // Acta Biotheor. 2003. V. 51. P. 35-41.
Roberts D.W. Ordination on the basis of fuzzy set theory // Vegetatio. 1986. V. 66. P. 123-131.
Roberts D.W. Statistical analysis of multidimensional fuzzy set ordinations // Ecology. 2008. V. 89, № 5. P. 1246-1260.
Ross G.J, Tasoulis D.K., Adams N.M. Nonparametric Monitoring of Data Streams for Changes in Location and Scale // Technometrics. 2011. V. 53, № 4. P. 379-389.
Rubinstein R.Y., Kroese D.P. Simulation and the Monte Carlo Method. Hoboken (NJ): Wiley & Sons, 2003. 336 p.
Saigo H. Comparing Four Bootstrap Methods for Stratified Three-Stage Sampling // Journal of Official Statistics. 2010. V. 26, №.1. P. 193-207.
Schaffer W. M. Stretching and folding in lynx fur returns: evidence for astrange attractor in nature? //Am. Nat. 1984. V. 124, № 6. P. 798-820.
Scherer R., Schaarschmidt F. Simultaneous confidence intervals for comparing biodiversity indices estimated from overdispersed count data // Biometrical Journal. 2013. V.55. P. 246–263.

Schluter D. A variance test for detecting species associations, with some example applications // Ecology. 1984. V. 65. P. 998-1005.
Seber G.A.F. Multivariate Observations. N.Y.: Wiley, 2004. 686 p.
Seefeld K., Linder E. Statistics Using R with Biological Examples. Durham (NH): University Press, 2007. URL: http://cran.r-project.org/doc/contrib/Seefeld_StatsRBio.pdf.
Shimodaira H. An approximately unbiased test of phylogenetic tree selection // Syst. Biol. 2002. V. 51. P. 492-508.
Singer J.D., Willett J.B. Applied Longitudinal Data Analysis: Modeling Change and Event Occurrence. N.Y.: Oxford University Press, 2003. 672 p.
Simon J. L. Resampling: the new statistics. Arlington (VA): Resampling Stats, 1997. 209 p.
Smouse P., Long J., Sokal R. Multiple Regression and Correlation Extensions of the Mantel Test of Matrix Correspondence // Systematic Zoology. 1986. V. 35, No. 4. P. 627-632
Soetaert K., Herman P. A Practical Guide to Ecological Modelling – Using R as a Simulation Platform. N.Y.: Springer, 2009. 372 p.
Sokal R.R., Rohlf F.J. Biometry: The principles and practice of statistics in biological research. N.Y.: Freeman, 1981. 887 p.
Stone L., Roberts A. The checkerboard score and species distributions // Oecologia. 1990. V. 85. P. 74-79.
Teetor P. 25 Recipes for Getting Started with R. Sebastopol (CA): O'Reilly Media, 2011. 62 p.
ter Braak C.J. Canonical correspondence analysis: a new eigenvector technique for multivariate direct gradient analysis // Ecology. 1986. V. 67. P. 1167-1179.
ter Braak C.J. Principal components biplots and alpha and beta diversity // Ecology. 1983. V. 64. P. 454-462.
Tichý L. JUICE, software for vegetation classification // J. Veg. Sci. 2002. V. 13. P. 451-453.
Tilman D. Functional diversity // Encyclopedia of Biodiversity (ed. Levin S.A.). San Diego: Academic Press, 2001. P. 109-120.
Tukey J.W. Bias and confidence in not quite large samples // Ann. Math. Statist. 1958. V. 29. P. 614.
Venables W.N., Ripley B.D. Modern Applied Statistics with S. N.Y.: Springer, 2003. 496 p.
Warwick R.M., Clarke K.R. Relearning the ABC: taxonomic changes and abundance/biomass relationships in disturbed benthic communities // Mar. Biol. 1994. V. 118, № 4. P. 739-744.
Warwick R.M., Clarke K.R. New 'biodiversity' measures reveal a decrease in taxonomic distinctness with increasing stress // Mar. Ecol. Prog. 1995. Ser. 129. P. 301-305.
Wehrens R. Chemometrics with R. N.Y.: Springer, 2011. 285 p.
Wood S. N. Generalized Additive Models: An Introduction with R. London: Chapman and Hall/CRC Press, 2006. 391 p.
Wright D.H., Patterson B.D., Mikkelson G.M. et al. A comparative analysis of nested subset patterns of species composition // Oecologia. 1998. V. 113. P. 1-20.
Zadeh L.A. Fuzzy sets // Information and Control. 1965. V. 8. P. 338-353.
Zieffler A.S., Harring J.R., Long J.D. Comparing Groups. Randomization and Bootstrap Methods Using R. Hoboken (NJ): Wiley & Sons Inc., 2011. 298 p.
Zuur A. F., Ieno E.N., Walker N. et al. Mixed Effects, Models and Extensions in Ecology with R. Berlin: Springer Sci., 2009. 574 p.
Zuur A.F., Fryer R.J., Jolliffe I.T. et al. Estimating common trends in multivariate time series using dynamic factor analysis // Environmetrics, 2003. V. 7. P. 665-685.

ПРИЛОЖЕНИЕ 1

Указатель использованных примеров и их краткое описание

1. Данные мониторинга экосистемы Куйбышевского водохранилища (1957-1988 гг).



Комплексные исследования института экологии Волжского бассейна на станциях наблюдения, расположенных в характерных точках акватории водоема, начиная от гг. Чебоксары и Чистополь (р. Кама) до плотины Волжской ГЭС у г. Тольятти. В разные годы число точек наблюдений колебалось от 10 до 53.

Экспедиционные рейсы осуществлялись в период открытой воды ежемесячно с мая по октябрь. На каждой станции со стандартных горизонтов отбирались пробы воды для учета численности и биомассы различных таксономических групп фитопланктона, зоопланктона и бактериопланктона, а также пробы грунта для исследования макрозообентоса.

Одновременно определялись основные физико-химические показатели качества воды и содержание биогенных и органических веществ. Данные мониторинга дополнялись информацией Госкомгидромета СССР (региональные гидрологические условия, температурный и радиационный режим, ветровая нагрузка и др.). Всего массив накопленных данных содержит более 94 000 измерений.

Избранные публикации:

Меншуткин В.В., Паутова В.Н., Номоконова и др. Статистические связи в экосистеме Куйбышевского водохранилища // Гидробиол. журн. 1998. Т. 34. № 5. С. 94-103.

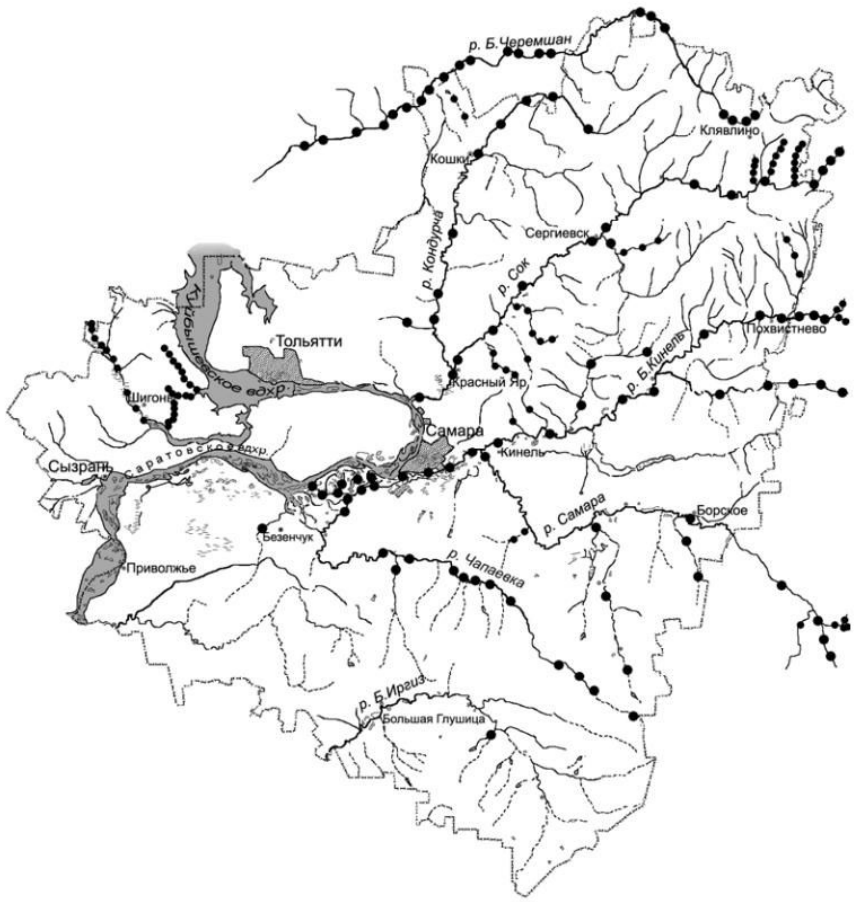
Шитиков В.К., Выхристюк Л.А., Паутова В.Н., Зинченко Т.Д. Многофакторное экологическое районирование Куйбышевского водохранилища // Водные ресурсы. 2007. Т. 34, № 4. С. 481-489.

Куйбышевское водохранилище (научно-информационный справочник) / Ред. Розенберг Г.С., Выхристюк Л.А. Тольятти: ИЭВБ РАН, 2008. 123 с.

Фрагменты базы данных использовались для иллюстрации примеров в разделах 2.3, 2.4, 4.2, 4.8, 6.9 и на протяжении всей главы 7.

2. Макрозообентос малых рек Средней Волги. Гидробиологическая съемка донных сообществ проводилась на 40 средних и малых реках бассейна Куйбышевского и Саратовского водохранилищ, протекающих в основном на территории Самарской обл.

В течении вегетативных сезонов 1986-2013 гг. на 228 створах рек было взято 557 гидробиологических проб, в которых было обнаружено 580 видов бентосных организмов. В ходе мониторинга водотоков оценивались также гидрологические и гидрохимические показатели.



В качестве отдельного модельного объекта рассматривалась средняя равнинная река Сок (375 км) вместе с ее притоком р. Байтуган (22 км). Здесь на 22 станциях наблюдения по результатам 147 гидробиологических проб в исходные таблицы было включено 375 видов и таксономических групп бентоса.

Исследования проводились лабораторией малых рек института экологии Волжского бассейна (рук. д.б.н. Зинченко Т.Д.). Избранные публикации:

Зинченко Т.Д. Эколого-фаунистическая характеристика хирономид (Diptera, Chironomidae) малых рек бассейна Средней и Нижней Волги (атлас). Тольятти: Кассандра, 2011. 258 с.

Особенности пресноводных экосистем малых рек Волжского бассейна / Ред. Розенберг Г.С., Зинченко Т.Д. Тольятти: Кассандра, 2011. 322 с.

Фрагменты базы данных использовались для иллюстрации примеров в разделах 1.4, 1.6, 2.3, 2.7, 2.8, 2.9, 3.1, 3.3, 3.6, 4.3, 4.8, 5.1, 5.2, 5.4, 5.6, 5.7, 6.2 и 6.3.

3. Геоботанические описания травянистого покрова в дельте р. Волга. В течение 1979-1981 гг. на 159 участках изучался состав травянистых растений (тростника, рагозы, пырея, череды, алтея и др.) и оценивалась в г/м^2 площади их суммарная надземная биомасса. По результатам химического анализа водной вытяжки определялся ионный состав почв: хлориды, сульфаты, бикарбонаты, кальций, натрий и магний. Для каждого участка измерялась средняя высота над уровнем межени, которая имеет сильную корреляцию со степенью увлажненности грунта.

Анализ собранных материалов является частью комплексных исследований растительности Волжского бассейна, проводимых лабораторией фитоценологии института экологии Волжского бассейна (рук. д.б.н. Голуб В.Б.). Ссылка на публикацию:

Голуб В.Б., Пилипенко В.Н., Лосев Г.А. Результаты оценки совместного влияния влажности и засоления почны на встречаемость растений дельты Волги и величину их надземной биомассы. // Биологические науки, 1986. № 7. С. 93-98.

Данные использовались для иллюстрации примеров в разделах 4.5, 4.6, 5.3, 5.5, 6.2, 6.3 и 6.5.

4. Морфометрические данные популяций полевки красной в районе оз. Байкал. Особи красной полевки (*Clethrionomys rutilus*) отлавливались в ходе экспедиционных исследований 1979-1983 гг. в темнохвойной тайге северных предгорий хребта Хамар-Дабан и в байкальской котловине на различном расстоянии (5÷100 км) от Байкальского целлюлозно-бумажного комбината (БЦБК). Предметом морфометрических исследований явились массо-размерные характеристики животных в целом и отдельных внутренних органов, а также промеры черепных элементов и координаты характерных точек мандибул. Всего с разной степенью детализации было изучено 1230 особей.

Материал любезно предоставлен проф. Коросовым А.В. (Петрозаводский государственный университет) и нашел отражение в публикациях, представленных на сайте http://korosov.narod.ru :

Коросов А. В., Демидович А. П. Оценка состояния населения мелких млекопитающих в зоне действия Байкальского целлюлозного комбината // Региональный мониторинг состояния озера Байкал. Л.: Гидрометеоиздат, 1987. С. 212-220.

Коросов А. В. Метод теневой остеометрии // Зоол. журн. 1988. Т. 68. Вып. 3. С. 121-125.

Фрагменты данных использовались для иллюстрации примеров в разделах 4.7, 6.6 и 6.7.

5. *Морфологические признаки и свойства ядовитого секрета популяций гадюки обыкновенной*. На территории европейской части России встречаются два подвида обыкновенной гадюки *Vipera berus* – номинативный *V. b. berus* и гадюка Никольского *V. b. Nikolskii.* Отлов животных проводился в Пермском крае (пос. Чепец и пос. Березовая Старица), а также в различных местообитаниях Вологодской, Пензенской, Липецкой, Саратовской и Самарской обл. Измерялись следующие морфологические признаки: количество брюшных щитков (*Ventr.*), количество пар подхвостовых (*S.cd.*) щитков, количество рядов чешуй вокруг середины тела (*Sq.*), количество верхнегубных (*Lab.*) и нижнегубных (*S.lab.*) щитков, количество чешуй вокруг глаза (*C.oc.*), количество мелких щитков между глазом и верхнегубными щитками (*S.oc.*), количество скуловых щитков (*Lor.*), количество интеркантальных чешуй (*Ic.*), число парафронтальных щитков (*Pf.*). В обобранном ядовитом секрете в лабораторных условиях определяли протеолитическую активность яда (ПА) и оксидазы *L*-аминокислот (*L*-АМО). Объем выборок – 100 животных, из них, 36 самцов и 64 самки.

Размерные показатели ящерицы живородящей и прыткой. Анализировались массо-размерные характеристики ящериц *Zootoca vivipara*, отловленных в различных районах Пермского края и Пензенской обл. Объем выборок – 92 самки, имеющие новорожденных животных и 86 взрослых самцов. Другая выборка ящерицы прыткой *Lacerta agilis* состояла из 213 особей, отловленных в окрестностях Тольятти и с. Мордово.

Сбор материала проводился лабораторией герпетологии института экологии Волжского бассейна (рук. к.б.н. Маленев А.Л., к.б.н. Епланова Г.В.).

Избранные публикации:

Маленев А.Л., Зайцева О.В., Бакиев А.Г., Зиненко А.И. Обыкновенная гадюка на границе речных бассейнов Волги и Дона: особенности морфологии змей и свойств ядовитого секрета в популяции из Пензенской области // Современная герпетология. 2010. Т. 10, вып. 3 / 4. С. 115–120.

Четанов Н.А., Епланова Г.В. Статистический анализ флуктуирующей асимметрии билатеральных признаков живородящей ящерицы *Zootoca vivipara* // Известия Самарского научного центра РАН. 2011. Т. 13, №1. С. 144 - 152

Фрагменты данных использовались для иллюстрации примеров в разделах 1.4, 1.5, 2.5, 2.6, 3.4, 6.8 и 7.7.

6. Состав липидов и жирных кислот ульвы в малых реках Приэльтонья. Образцы зеленой водоросли *Ulva intestinalis* отбирались в период с 2006 по 2010 гг. на 10 станциях рек Хара, Солянка, Сморогда, Чернавка, Ланцуг, питающих оз. Эльтон и имеющих различный уровень минерализации воды. Анализ липидного состава проводился лабораторией экологической биохимии института экологии Волжского бассейна (рук. д.б.н. Розенцвет О.А.). Ссылка на публикацию:

Розенцвет О.А., Нестеров В.Н., Богданова Е.С.. Влияние абиотических факторов на состав липидов *Ulva intestinalis* (L.) Link (Chlorophyta) в малых реках бассейна оз. Эльтон Прикаспийской низменности. // Биология внутренних вод, 2012. № 2. С. 214-221.

Фрагмент данных использовался для иллюстрации примера к разделу 2.6.

7. Полиморфизм локальных популяций улиток Крыма. В четырех физико-географических районах Крыма был проведен сбор 3115 экземпляров моллюсков *Helix albescens.* В лабораторных условиях был проанализирован характер опоясанности раковин и были рассчитаны частоты встречаемости отдельных морф.

Для иллюстрации примера к разделу 3.2 с согласия авторов использован материал статьи:

Крамаренко С.С., Леонов С.В. Фенетическая структура крымских популяций наземного моллюска *Helix albescens* (Gastropoda, Pulmonata, Helicidae) // Экология. 2011. № 2. С. 153-160.

Пространственное распределение популяций наземных моллюсков. Сбор материала осуществлялся в 2011 году в пределах стационара Днепропетровского национального университета им. О. Гончара. Пробные площадки исследуемого участка располагались вдоль 8 линий по 20 квадратов в каждой линии на расстоянии 1,5 м между их центрами (как по вертикали, так и по горизонтали). Улитки видов *Brephulopsis cylindrica* и *Monacha carthusiana* собирались в пределах квадрата 0,5×0,5 м и подсчитывалось число особей различных размерно-возрастных групп. Для иллюстрации примера к разделу 7.5 с согласия авторов использован материал статьи:

Винарский М.В., Крамаренко С.С., Лазуткина Е.А. и др. Статистические методы в изучении континентальных моллюсков // Статистические методы анализа в биологии и медицине. Омск: Вариант, 2012. С. 5-94.

8. Видовой состав рыбного населения Чебоксарского водохранилища. В период 1996-2006 гг. с использованием неводов с ячеей 6 мм на 67 станциях пяти участков водохранилища было отловлено 26596 особей 34 видов рыб. Были рассчитаны частоты встречаемости отдельных видов на каждом из участков.

Для иллюстрации примера к разделу 3.2 с согласия авторов использован материал статьи:

Минин А.Е., Постнов Д.И., Логинов В.В., Якимов В.Н. К вопросу о статистическом анализе пространственной структуры рыбного населения прибрежья Чебоксарского водохранилища по данным неводных съемок // Известия Калининградского государственного технического университета. 2011. Т. 22. С. 159-166.

9. Высотное распределение фауны дождевых червей Северного Кавказа. Сбор дождевых червей проведился в течение полевых сезонов 2002–2010 гг. на подветренных склонах Скалистого, Бокового и Главного Кавказского хребтов. Общий объем исследованного материала – более 7000 экз. из 149 биотопов разных высотных поясов Терского и Эльбрусского вариантов поясности, граница между которыми проходит по линии Дыхтау-Каракая в нижнем течении р. Баксан.

Материал любезно предоставлен к.б.н. Рапопорт И.Б. (Институт экологии горных территорий КБНЦ РАН). Ссылка на публикацию:

Рапопорт И. Б. Высотное распределение дождевых червей (Oligochaeta, Lumbricidae) в центральной части Северного Кавказа. // Зоологический журнал, 2013. Т. 92, № 1. С. 3-10.

Фрагмент данных использовался для иллюстрации примера к разделу 3.5.

Статистическая среда R и ее использование для обработки данных

Предуведомление

Подробно комментируемые скрипты, позволяющие выполнить расчеты по примерам, представленным в основных разделах, а также файлы с исходными данными для выполнения этих скриптов могут быть загружены из ресурса Интернет: http://www.ievbras.ru/ecostat/Kiril/Article/A32/Data.zip

Основные Интернет-ссылки

Система статистических расчётов R является свободно распространяемой программной средой с открытым исходным кодом, развиваемая в рамках проекта GNU. С основного сайта проекта (http://cran.r-project.org) можно бесплатно загрузить бинарные сборки для всех основных поддерживаемых платформ (Windows, UNIX и MacOS), большое число специализированных пакетов, а также, при желании, исходные тексты программных модулей. Система оснащена языком программирования высокого уровня для статистической обработки данных и работы с графикой.

Приведем некоторые русскоязычные сайты, поддерживаемые группами энтузиастов, где приводятся переводы документации по R, методические материалы, ведется обсуждение и обмен идеями, а также оказывается помощь новичкам в решении их проблем:

- http://r-analytics.blogspot.com/ - блог «Анализ и визуализация данных» С.Мастицкого;
- http://www.linuxformat.ru/archive.phtml - архив журнала Lunix Format (см. 2008 г №№ 1-12, 2010 г. №№ 2-5);
- http://r-statistics.livejournal.com/ - разделы в «Живом журнале»;
- http://www.inp.nsk.su/~baldin/DataAnalysis/index.html - материалы Е.Балдина;
- http://herba.msu.ru/shipunov/school/sch-ru.htm - материалы А.Шипунова;
- http://voliadis.ru/R-intro - материалы Р. Габидуллина;
- http://molbiol.ru/forums/index.php?showtopic=102724&st=600 – раздел в форуме «Биофизика и матметоды в биологии»;
- http://donbas-socproject.blogspot.com/search/label/учебник по R - блог «странного ученого» М.Касьянчука;
- http://chetvericov.ru/tag/r/#.UBu9waDB9K0 – блог «аспиранта-психолога» А. Четверикова;
- http://alexwin1961.livejournal.com/tag/r - блог А.Виноградова;
- http://vk.com/club8142131 - открытая группа ВКонтакте и т.д.,

а также наиболее интересные сайты, связанные с решением экологических задач:

- http://ecology.msu.montana.edu/labdsv/R/ - лабораторные работы по статистике в R для студентов экологического факультета университета в Монтане;
- http://www.bio.umontreal.ca/legendre/indexEn.html – страница П.Лежандра, проф. Монреальского университета, содержащая много скриптов и функций R.

Краткое руководство по установке и настройке среды

1. Скачать и установить базовую комплектацию статистической среды R; это можно сделать с основного сайта проекта http://cran.r-project.org или русского «зеркала» http://cran.gis-lab.info – всего около 40 Мб.

2. Выделить рабочий каталог для хранения скриптов, исходных данных и результатов. Удобно прописать путь к нему в системных файлах R, для чего добавить текстовым редактором в файл C:\Program Files\R\R-3.00.2\etc\Rprofile.site строчку
setwd("D:/R/Process/Resampling") # или любой иной каталог на Вашем компьютере. Весьма нежелательно использовать в названии рабочего каталога символы кириллицы.

3. Работать с системой R можно в командном окне R Console, что следует только приветствовать. Существуют, тем не менее, различные надстройки, позволяющие

проводить вычисления в среде, используя традиционную систему меню, например, R Commander, которую можно установить, выполнив команду загрузки пакета install.packages ("Rcmdr"). Это позволит вам осуществлять обработку данных (вывод графиков и расчет всех основных статистических критериев) в рамках регрессионного и дисперсионного анализа, не будучи знакомым с синтаксисом языка R. Познакомиться с более продвинутыми средствами интерфейса запуска и отладки скриптов можно, например, на сайте http://bioinformatics.ru/Data-Analysis/R-ide.html .

4. Скачать, распаковать Data.zip и копировать в рабочий каталог комплект скриптов и данных к отдельным главам книги (см. *Предуведомление*).

Основные конструкции языка среды R

Как и в любой вычислительной среде, основными компонентами R являются данные и операторы их обработки. Базовым объектом данных является вектор – проиндексированная последовательность числовых, символьных, логических, факториальных или иных специальных величин. Частными случаями векторов являются простая переменная, матрица (matrix – связанная совокупность однородных векторов), таблица данных (data.frame – матрица со столбцами разного типа) и список (list – иерархическая структура из векторов, матриц и т.д.).

Операторы обработки данных по синтаксису мало чем отличаются от традиционных языков программирования. Важнейшим отличием является громадное количество доступных функций, поэтому для уверенной работы в R нужна, прежде всего, хорошая память самого пользователя. Ниже приведены некоторые примеры операторов:

Операторы языка R	Проводимые действия
2*2 ; a <- 1/0	Делаем любые вычисления; переменной *a* присвоим ∞ (Inf)
a = factorial(10)/sin(2*pi)	Используем различные функции и встроенное число π
v = c(8.12,0,-64)^2	Создаем вектор *v* из квадратов трех числовых значений
seq(-5, 5, by=.2) -> v length(v)	Помещаем в вектор *v* последовательность чисел от -5 до 5 с шагом 0.2. Узнаем, что при этом получили 51 значение
v = scan("a.txt")	Загружаем в *v* данные из внешнего файла a.txt
sum(v) ; mean(v) ; median(v) var(v) ; sd(v)	Получаем сумму, среднее, медиану, дисперсию и стандартное отклонение членов совокупности *v*
v = rnorm(n=1000,mean=0,sd=1) hist(v) ; boxplot(v) ; qqplot(v) plot(density(v),col="red", lwd=2)	Формируем вектор 1000 случайных чисел из нормального распределения с нулевым средним и единичной дисперсией, рисуем гистограмму и графики плотности распределения
s <- c(LETTERS[7:12], letters[1:7]) sr <- sample(s) ; ss <- sr[order(sr)]	Создаем вектор из прописных и строчных букв от ‘a’ до ‘L’, случайно их перемешиваем, затем сортируем по алфавиту
f <- function (x, a) (2*x-a)^2 xmin <- optimize(f, c(0, 1), a = 0.35)	Находим минимум функции $(2x - 0.35)^2$ на интервале (0, 1)
A <- read.table(“As.txt”, header=F) A[is.na(A)] <- 0 ; n <-ncol(A) colnames(a) = c("sex","X","Y”)	Загружаем в таблицу *A* данные из внешнего файла As.txt, меняем пропущенные значения на 0, задаем имена столбцов
plot(AX, AY, pch=A$sex) table(A$sex) ; attach(A)	Строим «облако» точек зависимости Y = f(X) Подсчитываем число значений каждого класса sex
A<-transform(A,sexf= as.factor(sex)) summary(aov(X ~ sexf, data=A))	Добавляем в таблицу *A* столбец с фактором и выполняем дисперсионный анализ
m <- lm(Y~X+sexf) ; summary(m) predict(m, interval="confidence")	Получаем регрессионную модель зависимости Y= f(X, sex) и рассчитываем модельные значения

Как и в любом языке программирования, в процедурах R широко используются создание собственных функций, условный оператор if (условие) {выражение}, оператор цикла for (i in 1:n) {выражение} и т.д. Существуют развитые средства импорта исходных данных из внешних файлов различных форматов (.txt, .xls, .dbf, mdb и т.д.) или экспорта результатов в файлы .pdf, .doc и др.

При работе с использованием командного окна R Console мы рекомендуем во всех случаях создавать файл со скриптом (т.е. последовательностью операторов языка R, выполняющей определенные действия). Если этот файл существует, его лучше всего поместить в рабочий каталог, а если нет – создать там же новый текстовый файл с любым именем (но, для определенности, лучше с расширением *.r). Если после запуска R выбрать пункт меню "Файл > Открыть скрипт", то этот файл откроется в окне "Редактор R". Далее выделяется любой осмысленный фрагмент подготовленного скрипта (от имени одной переменной до всего содержимого) и осуществляется запуск этого блока на выполнение. Это можно сделать четырьмя возможными способами: (из основного и контекстного меню, кнопкой на панели инструментов и комбинацией клавиш Ctrl+R). Результаты, появившиеся в окне "Консоль", можно скопировать через буфер обмена в любой текстовый файл.

Установка используемых пакетов (package)

Стандартная инсталляция статистической среды R содержит базовую комплектацию из 17 пакетов (base, graphics, datasets, stats, splines, spatial, MASS, tcltk и др.), содержащих более 1600 различных функций. Этот список может быть существенно расширен в зависимости от профиля выполняемых работ: количество доступных пакетов уже превысило 2000 и их число неуклонно растет. Приведем неполный список пакетов, которые мы активно использовали в своих скриптах:

Пакеты R	Назначение
vegan, labdsv, simboot и FD	выполнение расчетов по экологии сообществ (меры сходства, разнообразия и вложенности, ординация и другие методы многомерного анализа)
picante	нуль-модели и анализ показателей в экологии и эволюции сообществ
ade4	анализ данных мониторинга окружающей среды
shapes	функции геометрической морфометрии
boot, bootstrap	функции различных процедур бутстрепа и «складного ножа»
lawstat	статистические критерии в биометрии
car	процедуры, связанные с прикладным регрессионным анализом
FWDselect, packfor,glmulti	выбор переменных в регрессионных моделях
lmodel2, lars	специальные виды регрессии (RMA, LARS, Lasso и др.)
lme4	построение моделей со смешанными эффектами
tree, rpart, randomForest, party, mvpart	построение иерархических деревьев регрессии
Hotelling	сравнение многомерных выборок
cluster	процедуры кластерного анализа
pvclust	бутстрепинг деревьев классификации
klaR, kknn, e1071	методы распознавания и классификации
nnet, kohonen	использование нейронных сетей и карт Кохонена
pastecs	анализ тренда временных рядов в экологии
forecast	прогнозирование временных рядов и модели сезонной динамики
spatstat	пространственное размещение точек, подбор моделей
spdep	пространственные зависимости: геостатистические методы и моделирование
rjags	связь с программой JAGS для получения марковских цепей Монте-Карло
xlsReadWrite	чтение и запись в файлы Excel
lattice	усовершенствованный графический пакет

Нет необходимости сразу устанавливать все перечисленные пакеты, если Вы не планируете рассмотреть все примеры: для каждого скрипта достаточно установить только те пакеты, которые объявляются функцией library(…). Для установки пакета достаточно в командном окне R Console выбрать пункт меню «Пакеты > Установить» или ввести, например, команду:
install.packages(c("vegan", "bootstrap", "boot", "lattice", "xlsReadWrite", "car"))
Пакеты можно скачивать, например, с русского «зеркала» http://cran.gis-lab.info.

Использование функций, сохраненных в файлах скриптов

Некоторые разработанные нами функции используются в скриптах, относящихся к разным разделам книги. Для их инициализации достаточно указать команду загрузки кода функций из файла, например:

source("print_rezult.r")

После этого становятся доступными следующие функции:

```
#-----------------------------------
#  Функция вывода результатов рандомизацационного теста произвольной статистики G
RandRes <- function (emp, sim, Nrand = 999, w.plot = 0) {
# Параметры: emp - эмпирическое значение  Gobs;
#                  sim - вектор рандомизированных значений  Gran
# Nrand - число итераций рандомизации;   w.plot <> 0  - выводится окно с гистограммой
#    Подготовка таблицы для вывода результатов рандомизации
CI <-  as.matrix(rep(NA, 7)) ; ; colnames(CI) <- "Стат"
rownames(CI) <- c("Эмпир.знач.","Средн.Ранд","CIлев", "CIправ",
"P(ранд>эмп)","P(ранд<эмп)","P(|ранд|>|эмп|)")
  CI[1] <-  emp ;   CI[2] <- mean(sim)
#    Доверительные интервалы методом процентилей
  CI[3] <- quantile(sim, prob=0.025) ;   CI[4] <- quantile(sim, prob=0.975)
#   Проверка односторонних гипотез
CI[5] <- (sum(sim >= emp)+1) / (Nrand + 1) ; CI[6]  <- (sum(sim <= emp)+1) / (Nrand + 1)
#   Проверка двухсторонней гипотезы
CI[7]   <- (sum(abs(sim)- abs(emp) >= 0)+1) / (Nrand + 1)
#   Расчет доверительных интервалов
CI.l <- quantile(sim, prob=0.025) ; CI.u <- quantile(sim, prob=0.925)
# Вывод гистограммы при необходимости
if (w.plot == 1)  { plot(hist(sim), col="gray80") ; abline(v= emp, lty=3,col=2,lwd=2) }
return(t(CI))}
#-----------------------------------
# Функция, выполняющая вывод результатов бутстепирования
#  Параметры: boots - выборка со значениями показателя, полученная в ходе бутстрепа
#                   empar  - показатель, рассчитанный по эмпирической выборке
BootRes <- function(boots,empar, w.plot = 0) {
if (w.plot == 1)    hist(boots)  # вывод гистограммы
#  Определение смещения и квантиля t-распределения. Формирование таблицы с результатами.
   bias <- mean(boots) - empar; tc <- qt(0.975, length(boots)-1) ;
CI <-  as.matrix(rep(NA, 7)) ; colnames(CI) <- "Стат"
rownames(CI) <- c("Эмпир.знач","Смещ","Ошибка","CIлев t","CIправ t","CIлев P", "CIправ P")
 CI[1] <-  empar ;   CI[2] <- bias ;   CI[3] <- sqrt(var(boots))
#  Доверительные интервалы с использованием t-распределения
  CI[4] <- empar - bias - tc* CI[3] ;  CI[5] <- empar - bias + tc* CI[3]
#  Доверительные интервалы методом процентилей
  CI[6] <- quantile(boots, prob=0.025);   CI[7] <- quantile(boots, prob=0.975)
  return(t(CI))}
```

Некоторые файлы содержат коды функций, которые необходимы для выполнения расчетов в полном объеме, но их текст не приводится в приложениях к разделам. Вот их полный список:

- pareto.R – функции оценки параметров распределения Парето (раздел 1.5);
- Kendall_Theil_Regr.r – оценка параметров линейной робастной регрессии Кендалла-Тейла (раздел 3.5);
- species_response_curves.r – подбор функции распределения популяционной плотности по градиенту (раздел 3.6);
- similary.r и abundsim.R – функции расчета различных метрик сходства (раздел 5.1);
- coldiss.R – раскраска матрицы расстояний (раздел 5.2);
- multi.mantel.R - матричная регрессия Мантеля (раздел 5.3);
- anova.1way.R – однофакторный дисперсионный анализ с рандомизацией (раздел 5.5);
- Xtree.r - создание матрицы длин ветвей по кластерной дендрограмме (раздел 5.7);
- cleanplot.pca.R – функция отрисовки биплота PCA (раздел 6.1);
- uis.r – сегментация временного ряда на бестрендовые участки (раздел 7.1);
- DF.r – тест Дики-Фуллера (раздел 7.2).

Все эти файлы включены в архив (Data.zip) с дополнительными материалами к книге.

Владимир Кириллович Шитиков
Геннадий Самуилович Розенберг

Рандомизация и бутстреп: статистический анализ в биологии и экологии с использованием R

Редактор *О.Л. Носкова*
Верстка и оригинал-макет *В.К. Шитиков*

Издательство «Кассандра»
445061, г. Тольятти, ул. Индустриальная, д. 7
Тел/факс: (8482) 570-004

Подписано в печать с оригинал макета 13.11.2013 г.
Формат 60x84 1/16 Бумага офсетная. Печать офсетная.
Усл. печ. л. 22,0
Тираж 100 экз. Заказ № 1315

Отпечатано в типографии ООО «Кассандра»